IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

DECLARATION OF VISHWANATH R. IYER, Ph.D.
UNDER 37 C.F.R. § 1.132

I, VISHWANATH R. IYER, Ph.D., declare and state as follows:

1. I am an Assistant Professor in the Section of Molecular Genetics and Microbiology, Institute of Cellular and Molecular Biology, University of Texas at Austin, where my laboratory currently studies global transcriptional control in yeast, gene expression programs during human cell proliferation, and genome-wide transcription factor targets in yeast and human. Immediately prior to this position, I spent four years as a postdoctoral fellow in the laboratory of Patrick O. Brown at Stanford University studying the transcriptional programs of yeast and of human cells. My *curriculum vitae* is attached hereto as Exhibit A.

2. Beginning in Dr. Brown's laboratory, where I helped to develop the first whole genome arrays for yeast and early versions of highly representative cDNA arrays for human cells, and continuing to the present day, I have used microarray-based gene expression analysis as a principal approach in much of my research.

3. Representative publications describing this work include:

DeRisi J. *et al.*, "Exploring the metabolic and genetic control of gene expression on a genomic scale," *Science* 278:680-686 (1997);[1]

Marton *et al.*, "Drug target validation and identification of secondary drug target effects using DNA microarrays," Nature Med. 4:1293-1301 (1998);[2]

Iyer *et al.*, "The transcriptional program in the response of human fibroblasts to serum," Science 283:83-87 (1999);[3] and

Ross *et al.*, "Systematic variation in gene expression patterns in human cancer cell lines," *Nature Genetics* 24: 227-235 (2000).[4]

Two of the papers describe our use of microarray-based expression profiling to explore the metabolic reprogramming that occurs during major environmental changes, both in yeast (DeRisi *et al.*, during the shift from fermentation to respiration) and in human cells (Iyer *et al.*, human fibroblasts exposed to serum). One reference describes our use of expression profile analysis in drug target validation and identification of secondary drug effects (Marton *et al.*). And one describes our use of expression profiling as a molecular phenotyping tool to discriminate among human cancer cells (Ross *et al.*).

4. Whether used to elucidate basic physiological responses, to study primary and secondary drug effects, or to discriminate and classify human cancers, expression profiling

---

[1] Attached hereto as Exhibit B.

[2] Attached hereto as Exhibit C.

[3] Attached hereto as Exhibit D.

[4] Attached hereto as Exhibit E.

as we have practiced it relies for its power on comparison of ***patterns*** of expression.

5. For example, we have demonstrated that we can use the presence or absence of a characteristic drug "signature" pattern of altered gene expression in drug-treated cells to explore the mechanism of drug action, and to identify secondary effects that can signal potentially deleterious drug side effects. As another example, we have demonstrated that gene expression patterns can be used to classify human tumor cell lines. While it is of course advantageous to know the biological function of the encoded gene products in order to reach a better understanding of the cellular mechanisms underlying these results, these pattern-based analyses do not require knowledge of the biological function of the encoded proteins.

6. The resolution of the patterns used in such comparisons is determined by the number of genes detected: the greater the number of genes detected, the higher the resolution of the pattern. It goes without saying that higher resolution patterns are generally more useful in such comparisons than lower resolution patterns. With such higher resolutions comes a correspondingly higher degree of statistical confidence for distinguishing different patterns, as well as identifying similar ones.

7. Each gene included as a probe on a microarray provides a signal that is specific to the cognate transcript, at least to a first approximation.[5] Each new gene-specific

---

[5] In a more nuanced view, it is certainly possible for a probe to signal the presence of a variety of splice variants of a single gene,

(Continued...)

probe added to a microarray thus increases the number of genes detectable by the device, increasing the resolving power of the device. As I note above, higher resolution patterns are generally more useful in comparisons than lower resolution patterns. Accordingly, each new gene probe added to a microarray increases the usefulness of the device in gene expression profiling analyses. This proposition is so well-established as to be virtually an axiom in the art, and has been as long as I have been working in the field, and certainly since the time I embarked on the production of whole genome arrays in early 1996. Simply put, arrays with fewer gene-specific probes are inferior to arrays with more gene-specific probes.

8. For example, our ability to subdivide cancers into discriminable classes by expression profiling is limited by the resolution of the patterns produced. With more genes contributing to the expression patterns, we can potentially draw finer distinctions among the patterns, thus subdividing otherwise indistinguishable cancers into a greater number of classes; the greater the number of classes, the greater the likelihood that the cancers classified together will respond similarly to therapeutic intervention, permitting better individualization of therapy and, we hope, better treatment outcomes.

9. If a gene does not change expression in an experiment, or if a gene is not expressed and produces no

---

(…Continued)
without discriminating among them, and for a probe to signal the presence of a variety of allelic variants of a single gene, again without discriminating among them.

signal in an experiment, that is not to say that the probe lacks usefulness on the array; it only means that an insufficient number of conditions have been sampled to identify expression changes. In fact, an experiment showing that a gene is not expressed or that its expression level does not change can be equally informative. To provide maximum versatility as a research tool, the microarray should include -- and as a biologist I would want my microarray to include -- each newly identified gene as a probe.

10. I declare further that all statements made herein of my own knowledge are true and that all statements made on information and belief are believed to be true, and further that these statements were made with the knowledge that willful false statements and the like so made are punishable by fine or imprisonment, or both, under Section 1001 of Title 18 of the United States Code and may jeopardize the validity of any patent application in which this declaration is filed or any patent that issues thereon.

Vishwanath

VISHWANATH R. IYER, Ph.D.

October 20, 2003

Date

BEST AVAILABLE IMAGES

Defective images within this document are accurate representations of the original documents submitted by the applicant.

Defects in the images may include (but are not limited to):

- BLACK BORDERS
- TEXT CUT OFF AT TOP, BOTTOM OR SIDES
- FADED TEXT
- ILLEGIBLE TEXT
- SKEWED/SLANTED IMAGES
- COLORED PHOTOS
- BLACK OR VERY BLACK AND WHITE DARK PHOTOS
- GRAY SCALE DOCUMENTS

IMAGES ARE BEST AVAILABLE COPY.

As rescanning documents *will not* correct images, please do not report the images to the Image Problem Mailbox.

Vishwanath R. Iyer

Assistant Professor
Section of Molecular Genetics and Microbiology
Institute of Cellular and Molecular Biology
MBB 3.212A, University of Texas at Austin
Austin, TX 78712-0159
Phone: 512-232-7833
Fax: 512-232-3432
Email: vishy@mail.utexas.edu

Education/Training

Bombay University, Mumbai, India	B.Sc. (1987), Chemistry & Biochemistry
M. S. University of Baroda, Baroda, India	M.Sc. (1989), Biotechnology
Harvard University, Cambridge MA	Ph.D. (1996), Genetics
Stanford University, Stanford CA	Post-doctoral (1996-2000), Genomics

Research Experience

9/00-5/03 Assistant professor, Section of Molecular Genetics and Microbiology, University of Texas, Austin TX

- Global transcriptional control in yeast
- Gene expression programs during human cell proliferation
- Genome-wide transcription factor targets in yeast and human
- Collaborative microarray facility

5/96-8/00 Post-doctoral fellow Stanford University, Stanford CA
(Advisor: Dr. Patrick O. Brown)

- Yeast whole-genome ORF and intergenic microarrays
- Human cDNA microarrays for expression profiling

9/89-4/96 Graduate student Harvard University, Cambridge MA
(Advisor: Dr. Kevin Struhl)

- Yeast transcriptional regulation

Honours and Awards

Government of India Biotechnology Fellowship (1987-1989)
University Grants Commission Junior Research Fellowship (1989)
Stanford University/NHGRI Genome Training Grant (1996)

Invited Conference talks (selected)

Invited Lecturer, NEC-Princeton Lectures in Biophysics
Princeton, NJ (June 1998)
Plenary Session Speaker, HGM '99 (HUGO Human Genome Meeting)
Brisbane, Australia (April 1999)
Invited Speaker, Gordon Research Conference "Human Molecular Genetics"
Newport, RI (August 2001)

Invited Speaker, Nature Genetics "Oncogenomics 2002" Conference
Dublin, Ireland (May 2002)
Invited Speaker, "Pathology Bioinformatics" Symposium, University of Michigan, Ann Arbor, MI (November 2002)
Invited Speaker, "Systems Biology: Genomic Approaches to Transcriptional Regulation" Cold Spring Harbor Laboratory Meeting (March 2003)
Symposium co-Chair and Speaker "Functional Genomics" American Society for Biochemistry and Molecular Biology Meeting, San Diego, CA (April 2003)
Invited Speaker in Functional Genomics (Gene Networks) Symposium, International Congress of Genetics, Melbourne Australia July 6-11 2003
Invited Speaker "BioArrays Europe 2003"
Cambridge, UK (Sep/Oct 2003)

Departmental Seminars

Texas A&M University Genetics and Biochemistry & Biophysics Departments, October 24 2002
New York University School of Medicine, Department of Biochemistry, November 20 2002
UT Southwestern Medical Center, Human Genetics Seminar Series, May 5 2002
UCLA School of Medicine, Department of Human Genetics
June 2 2003
National Human Genome Research Institute
June 12 2003
Sanger Institute of the Wellcome Trust, Hinxton, UK
Sep 2003

Other Professional Activities

Reviewer for *Genome Biology*, *Genome Research*, *Nature Genetics*, *Science* (1998-2003)
Instructor, Cold Spring Harbor Summer Course "Making and using DNA Microarrays" (2000 - 2003)
Member, NIDDK Special Emphasis Review Panel ZDK1 (2001-2002)

Publications

1. Iyer V. & Struhl, K. (1995) Poly(dA:dT), a ubiquitous promoter element that stimulates transcription via its intrinsic DNA structure, *EMBO J.* 14: 2570-2579.

2. Iyer V. & Struhl, K. (1995) Mechanism of differential utilization of the his3 TR and TC TATA elements, *Mol. Cell. Biol.* 15: 7059-7066.

3. Iyer V. & Struhl K. (1996) Absolute mRNA levels and transcription initiation rates in *Saccharomyces cerevisiae*. *Proc. Natl. Acad. Sci*. (*USA*) 93:5208-5212.

4. DeRisi J. L., Iyer V. R. & Brown P. O. (1997) Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science* 278:680-686

5. Marton M. J., DeRisi J. L., Bennett H. A., Iyer V. R., Meyer M. R., Roberts C. J., Stoughton R., Burchard J., Slade D., Dai H., Bassett D. E. Jr., Hartwell L. H., Brown P. O. & Friend S. H. (1998) Drug target validation and identification of secondary drug target effects using DNA microarrays. *Nature Med.* 4:1293-1301

6. Lutfiyya L. L., Iyer V. R., DeRisi J., DeVit M. J., Brown P. O. & Johnston M. (1998) Characterization of three related glucose repressors and genes they regulate in *Saccharomyces cerevisiae*. *Genetics* 150:1377-1391

7. Spellman P. T., Sherlock G., Zhang M. Q., Iyer V. R., Anders K., Eisen M. B., Brown P. O., Botstein D. & Futcher B. (1998) Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Mol. Biol. Cell* 9:3273-3297

8. Iyer V. R., Eisen M. B., Ross D. T., Schuler G., Moore T., Lee J. C.,F., Trent J. M., Staudt L. M., Hudson Jr. J., Boguski M. S., Lashkari D., Shalon D., Botstein D. & Brown P. O. (1999) The transcriptional program in the response of human fibroblasts to serum. *Science* 283:83-87

9. DeRisi J. L. & Iyer V. R. (1999) Genomics and array technology. *Curr. Opin. Oncol.* 11:76-79

10. Ross D. T., Scherf U., Eisen M. B., Perou C. M., Spellman P., Iyer V. R., Rees C., Jeffrey S. S., Van de Rijn M., Waltham M., Pergamenschikov A., Lee J. C. F., Lashkari D., Shalon D., Myers T. G., Weinstein J. N., Botstein D., & Brown P. O. (2000) Systematic variation in gene expression patterns in human cancer cell lines. *Nature Genetics* 24: 227-235

11. Sudarsanam P., Iyer V. R., Brown P. O. & Winston F. (2000) Whole-genome expression analysis of *snf/swi* mutants of *S. cerevisiae*. *Proc. Natl. Acad. Sci .(USA)* 97: 3364-3369

12. Tran H. G., Steger D. J., Iyer V. R., & Johnson A. D. (2000) The chromo domain protein Chd1p from budding yeast is an ATP-dependent chromatin-modifying factor *EMBO J* 19: 2323-2331

13. Gross C., Kelleher M., Iyer V. R., Brown P. O., & Winge D. R.. (2000) Identification of the copper regulon in *Saccharomyces cerevisiae* by DNA microarrays. *J. Biol. Chem.* 275: 32310-32316

14. Reid J. L., Iyer V. R., Brown P. O. & Struhl K. (2000) Coordinate regulation of yeast ribosomal protein genes is associated with targeted recruitment of Esa1 histone acetylase. *Mol. Cell* 6: 1297–1307

15. Iyer V. R., Horak C., Scafe C. S., Botstein D., Snyder M. & Brown P. O. (2001) Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF *Nature* 409: 533-538

16. Miki R., Kadota K., Bono H., Mizuno Y., Tomaru Y., Carninci P., Itoh M., Shibata K., Kawai J., Konno H., Watanabe S., Sato K., Tokusumi Y., Kikuchi N., Ishii Y., Hamaguchi Y., Nishizuka I., Goto H., Nitanda H., Satomi S., Yoshiki A., Kusakabe M., DeRisi J.L., Eisen M.B., Iyer V.R., Brown P.O., Muramatsu M., Shimada H., Okazaki Y. & Hayashizaki Y. (2001) Delineating developmental and metabolic pathways in vivo by expression profiling using the RIKEN set of 18,816 full-length enriched mouse cDNA arrays *Proc. Natl. Acad. Sci. (USA)* 98: 2199-2204

17. Pollack J. R. & Iyer V.R. (2002) Characterizing the physical genome. *Nature Genetics* 32 suppl: 515-521

18. Iyer V. R. Microarray-based detection of DNA protein interactions: Chromatin Immunoprecipitation on Microarrays, in *DNA Microarrays: A Molecular Cloning Manual* (eds. Bowtell, D. & Sambrook, J.) 453-463 (Cold Spring Harbor Laboratory Press, 2003).
*(*not peer reviewed*)

19. Killion, P., Sherlock G. and Iyer V. R. (2003) The Longhorn Array Database, an open-source implementation of the Stanford Microarray Database *BMC Bioinformatics* 4: 32

20. Hahn J. S., Hu Z., Thiele D. J. & Iyer V. R. Genome-Wide Analysis of the Biology of Stress Responses Through Heat Shock Transcription Factor (submitted to *PNAS*)

21. Kim J. & Iyer V.R. The global role of TBP recruitment to promoters in mediating gene expression profiles (manuscript in preparation)

Current/Pending Research Support

U01 AA13518-01 Adron Harris (PI) 25% effort
9/28/01 - 9/27/06
NIH/NIAAA
"INIA: Microarray Core"
This proposal was a response to the Integrative Neuroscience Initiative on Alcoholism (INIA) RFA-AA-01-002. The overall goal is to support the use of microarray technology to define changes in gene expression that either predict or accompany excessive alcohol consumption.
Role: Co-investigator

003658-0223-2001 Iyer (PI) 16% effort
01/01/02 - 08/31/04
Texas Higher Education Coordinating Board (ARP)
"Microarray based global mapping of DNA-protein interactions at promoters in human cells"
This is a pilot project to map the in vivo interactions of transcription factors with human promoters
Role: PI

Information Technology Research 0325116 R. Mooney (PI) 9% effort
09/01/03 - 08/31/07
NSF
"Feedback from Multi-Source Data Mining to Experimentation for Gene Network Discovery"
Role: Co-investigator

1 R01 CA95548-01A2 (pending) Iyer (PI) 25% effort
12/1/03 – 11/30/08
NIH
"Analysis of genome-wide transcriptional control in yeast"
This is a project to identify stress responsive transcription factor targets in yeast through the use of DNA microarrays
Role: PI

Breast Cancer Idea Award (pending) Iyer (PI) 10% effort
1/1/04 – 12/31/06
US Army Medical Research and Materiel Command
"Genome-wide chromosomal targets of oncogenic transcription factors"
This is a project aimed at identifying direct chromosomal targets of c-myc and ER in human cells through the use of a novel sequence tag analysis method.
Role: PI

003658-0531-2003 (pending) Marcotte (PI) 8% effort
01/01/04 - 12/31/05
Texas Higher Education Coordinating Board (ATP)
"Cell arrays: A novel high–throughput platform for measuring gene function on a genomic scale"
This proposal is aimed at developing a novel microarray based platform for automated, high–throughput microscopic imaging of cells, allowing rapid and systematic evaluation of gene function.

Fischer-Vize, *Science* 270, 1828 (1995).
35. T. C. James and S. C. Elgin, *Mol. Cell Biol.* 6, 3862 (1986); R. Paro and D. S. Hogness, *Proc. Natl. Acad. Sci. U.S.A.* 88, 263 (1991); B. Tschiersch *et al.*, *EMBO J.* 13, 3822 (1994); M. T. Madireddi *et al.*, *Cell* 87, 75 (1996); D. G. Stokes, K. D. Tartof, R. P. Perry, *Proc. Natl. Acad. Sci. U.S.A.* 93, 7137 (1996).
36. P. M. Palosaari *et al.*, *J. Biol. Chem.* 266, 10750 (1991); A. Schmitz, K. H. Gartemann, J. Fiedler, E. Grund, R. Eichenlaub, *Appl. Environ. Microbiol.* 58, 4068 (1992); V. Sharma, K. Suvarna, R. Meganathan, M. E. Hudspeth, *J. Bacteriol.* 174, 5057 (1992); M. Kanazawa *et al.*, *Enzyme Protein* 47, 9 (1993); Z. L. Boynton, G. N. Bennet, F. B. Rudolph, *J. Bacteriol.* 178, 3015 (1996).
37. M. Ho *et al.*, *Cell* 77, 869 (1994).
38. W. Hendriks *et al.*, *J. Cell Biochem.* 59, 418 (1995).
39. We thank H. Skaletsky and F. Lewitter for help with

Exploring the Metabolic and Genetic Control of Gene Expression on a Genomic Scale

Joseph L. DeRisi, Vishwanath R. Iyer, Patrick O. Brown*

DNA microarrays containing virtually every gene of *Saccharomyces cerevisiae* were used to carry out a comprehensive investigation of the temporal program of gene expression accompanying the metabolic shift from fermentation to respiration. The expression profiles observed for genes with known metabolic functions pointed to features of the metab lic reprogramming that occur during the diauxic shift, and the expression patterns of many previously uncharacterized genes provided clues to their possible functions. The same DNA microarrays were also used to identify genes whose expression was affected by deletion of the transcriptional co-repressor *TUP1* or overexpression of the transcriptional activator *YAP1*. These results demonstrate the feasibility and utility of this approach to genomewide exploration of gene expression patterns.

The complete sequences of nearly a dozen microbial genomes are known, and in the next several years we expect to know the complete genome sequences of several metazoans, including the human genome. Defining the role of each gene in these genomes will be a formidable task, and understanding how the genome functions as a whole in the complex natural history of a living organism presents an even greater challenge.

Knowing when and where a gene is expressed often provides a strong clue as to its biological role. Conversely, the pattern of genes expressed in a cell can provide detailed information about its state. Although regulation of protein abundance in a cell is by no means accomplished solely by regulation of mRNA, virtually all differences in cell type or state are correlated with changes in the mRNA levels of many genes. This is fortuitous because the only specific reagent required to measure the abundance of the mRNA for a specific gene is a cDNA sequence. DNA microarrays, consisting of thousands of individual gene sequences printed in a high-density array on a glass microscope slide (*1*, *2*), provide a practical and economical tool for studying gene expression on a very large scale (*3*–*6*).

Saccharomyces cerevisiae is an especially favorable organism in which to conduct a systematic investigation of gene expression. The genes are easy to recognize in the genome sequence, *cis* regulatory elements are generally compact and close to the transcription units, much is already known about its genetic regulatory mechanisms, and a powerful set of tools is available for its analysis.

A recurring cycle in the natural history of yeast involves a shift from anaerobic (fermentation) to aerobic (respiration) metabolism. Inoculation of yeast into a medium rich in sugar is followed by rapid growth fueled by fermentation, with the production of ethanol. When the fermentable sugar is exhausted, the yeast cells turn to ethanol as a carbon source for aerobic growth. This switch from anaerobic growth to aerobic respiration upon depletion of glucose, referred to as the diauxic shift, is correlated with widespread changes in the expression of genes involved in fundamental cellular processes such as carbon metabolism, protein synthesis, and carbohydrate storage (*7*). We used DNA microarrays to characterize the changes in gene expression that take place during this process for nearly the entire genome, and to investigate the genetic circuitry that regulates and executes this program.

Yeast open reading frames (ORFs) were amplified by the polymerase chain reaction (PCR), with a commercially available set of primer pairs (*8*). DNA microarrays, containing approximately 6400 distinct DNA sequences, were printed onto glass slides by using a simple robotic printing device (*9*). Cells from an exponentially growing culture of yeast were inoculated into fresh medium and grown at 30°C for 21 hours. After an initial 9 hours of growth, samples were harvested at seven successive 2-hour intervals, and mRNA was isolated (*10*). Fluorescently labeled cDNA was prepared by reverse transcription in the presence of Cy3(green)- or Cy5(red)-labeled deoxyuridine triphosphate (dUTP) (*11*) and then hybridized to the microarrays (*12*). To maximize the reliability with which changes in expression levels could be discerned, we labeled cDNA prepared from cells at each successive time point with Cy5, then mixed it with a Cy3-labeled "reference" cDNA sample prepared from cells harvested at the first interval after inoculation. In this experimental design, the relative fluorescence intensity measured for the Cy3 and Cy5 fluors at each array element provides a reliable measure of the relative abundance of the corresponding mRNA in the two cell populations (Fig. 1). Data from the series of seven samples (Fig. 2), consisting of more than 43,000 expression-ratio measurements, were organized into a database to facilitate efficient exploration and analysis of the results. This database is publicly available on the Internet (*13*).

During exponential growth in glucose-rich medium, the global pattern of gene expression was remarkably stable. Indeed, when gene expression patterns between the first two cell samples (harvested at a 2-hour interval) were compared, mRNA levels differed by a factor of 2 or more for only 19 genes (0.3%), and the largest of these differences was only 2.7-fold (*14*). However, as glucose was progressively depleted from the growth media during the course of the experiment, a marked change was seen in the global pattern of gene expression. mRNA levels for approximately 710 genes were induced by a factor of at least 2, and the mRNA levels for approximately 1030 genes declined by a factor of at least 2. Messenger RNA levels for 183 genes increased by a factor of at least 4, and mRNA levels for 203 genes diminished by a factor of at least 4. About half of these differentially expressed genes have no currently recognized function and are not yet named. Indeed, more than 400 of the differentially expressed genes have no apparent homology

Department of Biochemistry, Stanford University School of Medicine, Howard Hughes Medical Institute, Stanford, CA 94305-5428, USA.

*To whom correspondence should be addressed. E-mail: pbrown@cmgm.stanford.edu

to any gene whose function is known (*15*). The responses of these previously uncharacterized genes to the diauxic shift therefore provides the first small clue to their possible roles.

The global view of changes in expression f genes with known functions provides a vivid picture of the way in which the cell adapts to a changing environment. Figure 3 shows a portion of the yeast metabolic pathways involved in carbon and energy metabolism. Mapping the changes we observed in the mRNAs encoding each enzyme onto this framework allowed us to infer the redirection in the flow of metabolites through this system. We observed large inductions of the genes coding for the enzymes aldehyde dehydrogenase (*ALD2*) and acetyl–coenzyme A(CoA) synthase (*ACS1*), which function together to convert the products of alcohol dehydrogenase into acetyl-CoA, which in turn is used to fuel the tricarboxylic acid (TCA) cycle and the glyoxylate cycle. The concomitant shutdown of transcription of the genes encoding pyruvate decarboxylase and induction of pyruvate carboxylase rechannels pyruvate away from acetaldehyde, and instead to oxalacetate, where it can serve to supply the TCA cycle and gluconeogenesis. Induction of the pivotal genes *PCK1*, encoding phosphoenolpyruvate carboxykinase, and *FBP1*, encoding fructose 1,6-biphosphatase, switches the directions of two key irreversible steps in glycolysis, reversing the flow of metabolites along the reversible steps of the glycolytic pathway toward the essential biosynthetic precursor, glucose-6-phosphate. Induction of the genes coding for the trehalose synthase and glycogen synthase complexes promotes channeling of glucose-6-phosphate into these carbohydrate storage pathways.

Just as the changes in expression of genes encoding pivotal enzymes can provide insight into metabolic reprogramming, the behavior of large groups of functionally related genes can provide a broad view of the systematic way in which the yeast cell adapts to a changing environment (Fig. 4). Several classes of genes, such as cytochrome c–related genes and those involved in the TCA/glyoxylate cycle and carbohydrate storage, were coordinately induced by glucose exhaustion. In contrast, genes devoted to protein synthesis, including ribosomal proteins, tRNA synthetases, and translation, elongation, and initiation factors, exhibited a coordinated decrease in expression. More than 95% of ribosomal genes showed at least twofold decreases in expression during the diauxic shift (Fig. 4) (*13*). A noteworthy and illuminating exception was that the genes encoding mitochondrial ribosomal genes were generally induced rather than repressed after glucose limitation, highlighting the requirement for mitchondrial biogenesis (*13*). As more is learned about the functions of every gene in the yeast genome, the ability to gain insight into a cell's response to a changing environment through its global gene expression patterns will become increasingly powerful.

Several distinct temporal patterns of expression could be recognized, and sets of genes could be grouped on the basis of the similarities in their expression patterns. The characterized members of each of these groups also shared important similarities in their functions. Moreover, in most cases, common regulatory mechanisms could be inferred for sets of genes with similar expression profiles. For example, seven genes showed a late induction profile, with mRNA levels increasing by more than ninefold at the last timepoint but less than threefold at the preceding timepoint (Fig. 5B). All of these genes were known to be glucose-repressed, and five of the seven were previously noted t share a common upstream activating sequence (UAS), the carbon source response element (CSRE) (*16–20*). A search in the promoter regions of the remaining two genes, *ACR1* and *IDP2*, revealed that *ACR1*, a gene essential for *ACS1* activity, also possessed a consensus CSRE motif, but interestingly, *IDP2* did not. A search of the entire yeast genome sequence for the consensus CSRE motif revealed only four additional candidate genes, none of which showed a similar induction.

Examples from additional groups of genes that shared expression profiles are illustrated in Fig. 5, C through F. The sequences upstream of the named genes in Fig. 5C all contain stress response elements (STRE), and with the exception

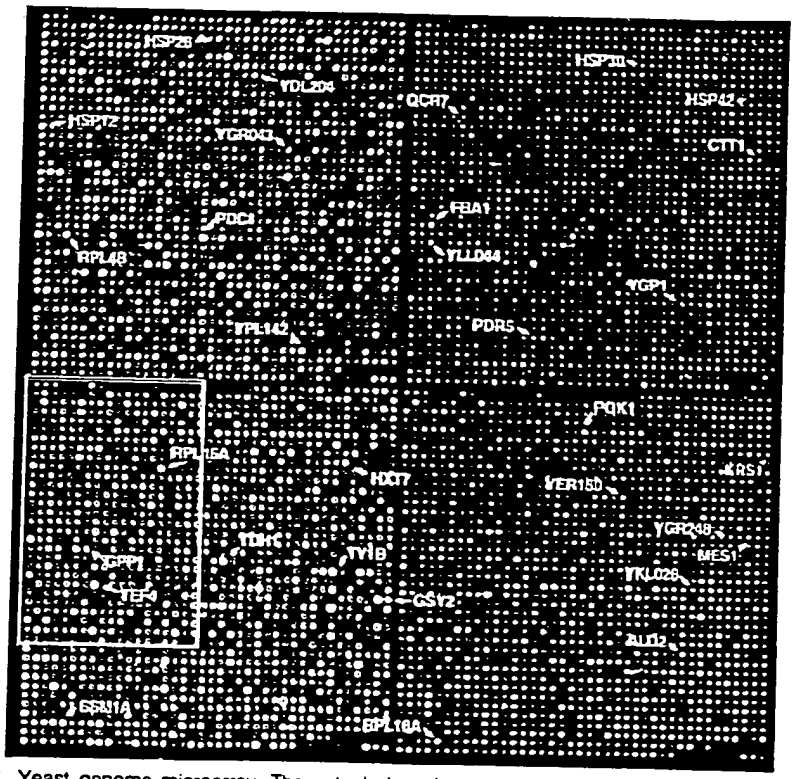

Fig. 1. Yeast genome microarray. The actual size of the microarray is 18 mm by 18 mm. The microarray was printed as described (*9*). This image was obtained with the same fluorescent scanning confocal microscope used to collect all the data we report (*49*). A fluorescently labeled cDNA probe was prepared from mRNA isolated from cells harvested shortly after inoculation (culture density of $<5 \times 10^6$ cells/ml and media glucose level of 19 g/liter) by reverse transcription in the presence of Cy3-dUTP. Similarly, a second probe was prepared from mRNA isolated from cells taken from the same culture 9.5 hours later (culture density of $\sim 2 \times 10^8$ cells/ml, with a glucose level of <0.2 g/liter) by reverse transcription in the presence of Cy5-dUTP. In this image, hybridization of the Cy3-dUTP–labeled cDNA (that is, mRNA expression at the initial timepoint) is represented as a green signal, and hybridization of Cy5-dUTP–labeled cDNA (that is, mRNA expression at 9.5 hours) is represented as a red signal. Thus, genes induced or repressed after the diauxic shift appear in this image as red and green spots, respectively. Genes expressed at roughly equal levels before and after the diauxic shift appear in this image as yellow spots.

of *HSP42*, have previ usly been shown t be contr lled at least in part by these elements (*21–24*). Inspection of the sequences upstream f *HSP42* and the two uncharacterized genes shown in Fig. 5C, YKL026c, a hypothetical protein with similarity to glutathione peroxidase, and YGR043c, a putative transaldolase, revealed that each of these genes also possess repeated upstream copies of the stress-responsive CCCCT motif. Of the 13 additional genes in the yeast genome that shared this expression profile [including *HSP30*, *ALD2*, *OM45*, and 10 uncharacterized ORFs (*25*)], nine contained one or more recognizable STRE sites in their upstream regions.

The heterotrimeric transcriptional activator complex *HAP2,3,4* has been shown to be responsible for induction of several genes important for respiration (*26–28*). This complex binds a degenerate consensus sequence known as the CCAAT box (*26*). Computer analysis, using the consensus sequence TNRYTGGB (*29*), has suggested that a large number of genes involved in respiration may be specific targets of *HAP2,3,4* (*30*). Indeed, a putative *HAP2,3,4* binding site could be found in the sequences upstream of each of the seven cytochrome c–related genes that showed the greatest magnitude of induction (Fig. 5D). Of 12 additional cytochrome c–related genes that were induced, *HAP2,3,4* binding sites were present in all but one. Significantly, we found that transcription of *HAP4* itself was induced nearly ninefold concomitant with the diauxic shift.

Control of ribosomal protein biogenesis is mainly exerted at the transcriptional level, through the presence of a common upstream-activating element (UAS_{rpg}) that is recognized by the Rap1 DNA-binding protein (*31*, *32*). The expression profiles of seven ribosomal proteins are shown in Fig. 5F. A search of the sequences upstream of all seven genes revealed consensus Rap1-binding motifs (*33*). It has been suggested that declining Rap1 levels in the cell during starvation may be responsible for the decline in ribosomal protein gene expression (*34*). Indeed, we observed that the abundance of *RAP1* mRNA diminished by 4.4-fold, at about the time of glucose exhaustion.

Of the 149 genes that encode known or putative transcription factors, only two, *HAP4* and *SIP4*, were induced by a factor of more than threefold at the diauxic shift. *SIP4* encodes a DNA-binding transcriptional activator that has been shown to interact with Snf1, the "master regulator" of glucose repression (*35*). The eightfold induction of *SIP4* upon depletion of glucose strongly suggests a role in the induction of downstream genes at the diauxic shift.

Although most of the transcriptional responses that we observed were not previously known, the responses of many genes during the diauxic shift have been described. Comparison of the results we obtained by DNA microarray hybridization with previously reported results therefore provided a strong test of the sensitivity and accuracy of this approach. The expression patterns we observed for previously characterized genes showed almost perfect concordance with previously published results (*36*). Moreover, the differential expression measurements obtained by DNA microarray hybridization were reproducible in duplicate experiments. For example, the remarkable changes in gene expression between cells harvested immediately after inoculation and immediately after the diauxic shift (the first and sixth intervals in this time series) were measured in duplicate, independent DNA microarray hybridizations. The correlation coefficient for two complete sets of expression ratio measurements was 0.87, and for more than 95% of the genes, the expression ratios measured in these duplicate experiments differed by less than a factor of 2. However, in a few cases, there were discrepancies between our results and previous results, pointing to technical limitations that will need to be addressed as DNA microarray technology advances (*37*, *38*). Despite the noted exceptions, the high concordance between the results we obtained in these experiments and those of previous studies provides confidence in the reliability and thoroughness of the survey.

The changes in gene expression during this diauxic shift are complex and involve integration of many kinds of information about the nutritional and metabolic state of the cell. The large number of genes whose expression is altered and the diversity of temporal expression profiles observed in this experiment highlight the challenge of understanding the underlying regulatory mechanisms. One approach to defining the contributions of individual regulatory genes to a complex program of this kind is to use DNA microarrays to identify genes whose expression is affected

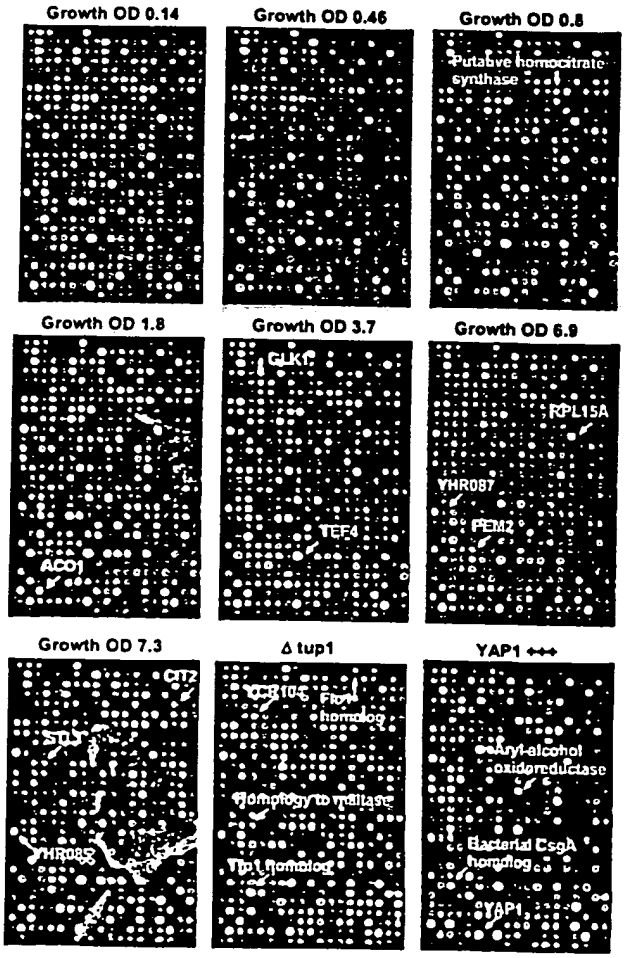

Fig. 2. The section of the array indicated by the gray box in Fig. 1 is shown for each of the experiments described here. Representative genes are labeled. In each of the arrays used to analyze gene expression during the diauxic shift, red spots represent genes that were induced relative to the initial timepoint, and green spots represent genes that were repressed relative to the initial timepoint. In the arrays used to analyze the effects of the *tup1*Δ mutation and *YAP1* overexpression, red spots represent genes whose expression was increased, and green spots represent genes whose expression was decreased by the genetic modification. Note that distinct sets of genes are induced and repressed in the different experiments. The complete images of each of these arrays can be viewed on the Internet (*13*). Cell density as measured by optical density (OD) at 600 nm was used to measure the growth of the culture.

by mutations in each putative regulatory gene. As a test of this strategy, we analyzed the genomewide changes in gene expression that result from deleti n of the *TUP1* gene. Transcriptional repression f many genes by glucose requires the DNA-binding repressor Mig1 and is mediated by recruiting the transcriptional co-repressors Tup1 and Cyc8/Ssn6 (39). Tup1 has also been implicated in repression of oxygen-regulated, mating-type–specific, and DNA-damage–inducible genes (40).

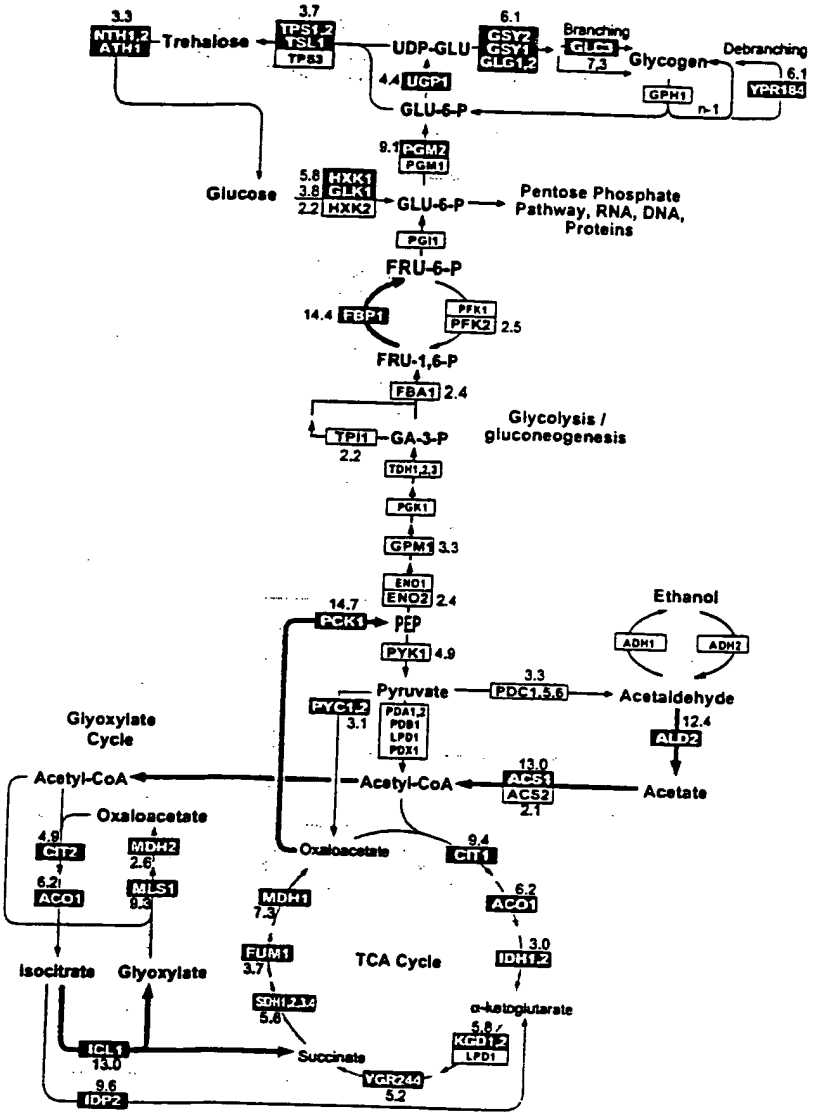

Fig. 3. Metabolic reprogramming inferred from global analysis of changes in gene expression. Only key metabolic intermediates are identified. The yeast genes encoding the enzymes that catalyze each step in this metabolic circuit are identified by name in the boxes. The genes encoding succinyl-CoA synthase and glycogen-debranching enzyme have not been explicitly identified, but the ORFs YGR244 and YPR184 show significant homology to known succinyl-CoA synthase and glycogen-debranching enzymes, respectively, and are therefore included in the corresponding steps in this figure. Red boxes with white lettering identify genes whose expression increases in the diauxic shift. Green boxes with dark green lettering identify genes whose expression diminishes in the diauxic shift. The magnitude of induction or repression is indicated for these genes. For multimeric enzyme complexes, such as succinate dehydrogenase, the indicated fold-induction represents an unweighted average of all the genes listed in the box. Black and white boxes indicate no significant differential expression (less than twofold). The direction of the arrows connecting reversible enzymatic steps indicate the direction of the flow of metabolic intermediates, inferred from the gene expression pattern, after the diauxic shift. Arrows representing steps catalyzed by genes whose expression was strongly induced are highlighted in red. The broad gray arrows represent major increases in the flow of metabolites after the diauxic shift, inferred from the indicated changes in gene expression.

Wild-type yeast cells and cells bearing a deletion f the *TUP1* gene (*tup1Δ*) were grown in parallel cultures in rich medium containing glucose as the carbon source. Messenger RNA was isolated from exponentially growing cells from the two populations and used to prepare cDNA labeled with Cy3 (green) and Cy5 (red), respectively (*11*). The labeled probes were mixed and simultaneously hybridized t the microarray. Red spots on the microarray therefore represented genes whose transcription was induced in the *tup1Δ* strain, and thus presumably repressed by Tup1 (*41*). A representative section of the microarray (Fig. 2, bottom middle panel) illustrates that the genes whose expression was affected by the *tup1Δ* mutation, were, in general, distinct from those induced upon glucose exhaustion [complete images of all the arrays shown in Fig. 2 are available on the Internet (*13*)]. Nevertheless, 34 (10%) of the genes that were induced by a factor of at least 2 after the diauxic shift were similarly induced by deletion of *TUP1*, suggesting that these genes may be subject to *TUP1*-mediated repression by glucose. For example, *SUC2*, the gene encoding invertase, and all five hexose transporter genes that were induced during the course of the diauxic shift were similarly induced, in duplicate experiments, by the deletion of *TUP1*.

The set of genes affected by Tup1 in this experiment also included α-glucosidases, the mating-type–specific genes *MFA1* and *MFA2*, and the DNA damage–inducible *RNR2* and *RNR4*, as well as genes involved in flocculation and many genes of unknown function. The hybridization signal corresponding to expression of *TUP1* itself was also severely reduced because of the (incomplete) deletion of the transcription unit in the *tup1Δ* strain, providing a positive control in the experiment (*42*).

Many of the transcriptional targets of Tup1 fell into sets of genes with related biochemical functions. For instance, although only about 3% of all yeast genes appeared to be *TUP1*-repressed by a factor of more than 2 in duplicate experiments under these conditions, 6 of the 13 genes that have been implicated in flocculation (*15*) showed a reproducible increase in expression of at least twofold when *TUP1* was deleted. Another group of related genes that appeared to be subject to *TUP1* repression encodes the serine-rich cell wall mannoproteins, such as Tip1 and Tir1/Srp1 which are induced by cold shock and other stresses (*43*), and similar, serine-poor proteins, the seripauperins (*44*). Messenger RNA levels for 23 of the 26 genes in this group were reproducibly elevated by at least 2.5-fold in the *tup1Δ*

strain, and 18 f these genes were induced by more than sevenfold when *TUP1* was deleted. In contrast, none of 83 genes that could be classified as putative regulators f the cell divisi n cycle were induced more than tw f ld by deletion of *TUP1*. Thus, despite the diversity of the regulatory systems that employ Tup1, most of the genes that it regulates under these conditions fall into a limited number of distinct functional classes.

Because the microarray allows us to monitor expression of nearly every gene in yeast, we can, in principle, use this approach to identify all the transcriptional targets of a regulatory protein like Tup1. It is important to note, however, that in any single experiment of this kind we can only recognize those target genes that are normally repressed (or induced) under the conditions of the experiment. For instance, the experiment described here analyzed a *MAT* α strain in which *MFA1* and *MFA2*, the genes encoding the a-factor mating pheromone precursor, are normally repressed. In the isogenic *tup1Δ* strain, these genes were inappropriately expressed, reflecting the role that Tup1 plays in their repression. Had we instead carried out this experiment with a *MATA* strain (in which expression of *MFA1* and *MFA2* is not repressed), it would not have been possible to conclude anything regarding the role of Tup1 in the repression of these genes. Conversely, we cannot distinguish indirect effects of the chronic absence of Tup1 in the mutant strain from effects directly attributable to its participation in repressing the transcription of a gene.

Another simple route to modulating the activity of a regulatory factor is to overexpress the gene that encodes it. *YAP1* encodes a DNA-binding transcription factor belonging to the b-zip class of DNA-binding proteins. Overexpression of *YAP1* in yeast confers increased resistance to hydrogen peroxide, o-phenanthroline, heavy metals, and osmotic stress (*45*). We analyzed differential gene expression between a wild-type strain bearing a control plasmid and a strain with a plasmid expressing *YAP1* under the control of the strong *GAL1-10* promoter, both grown in galactose (that is, a condition that induces *YAP1* overexpression). Complementary DNA from the control and *YAP1* overexpressing strains, labeled with Cy3 and Cy5, respectively, was prepared from mRNA isolated from the two strains and hybridized to the microarray. Thus, red spots on the array represent genes that were induced in the strain overexpressing *YAP1*.

Of the 17 genes whose mRNA levels increased by more than threefold when *YAP1* was overexpressed in this way, five bear homology to aryl-alcohol oxidoreductases (Fig. 2 and Table 1). An additional four of the genes in this set also belong to the general class of dehydrogenases/oxidoreductases. Very little is known about the role of aryl-alcohol oxidoreductases in *S. cerevisiae*, but these enzymes have been isolated from ligninolytic fungi, in which they participate in coupled redox reactions, oxidizing aromatic, and aliphatic unsaturated alcohols to aldehydes with the production of hydrogen peroxide (*46*, *47*). The fact that a remarkable fraction of the targets identified in this experiment belong to the same small, functional group of oxidoreductases suggests that these genes might play an important protective role during oxidative stress. Transcription of a small number of genes was reduced in the strain overexpressing Yap1. Interestingly, many of these genes encode sugar permeases or enzymes involved in inosit l metabolism.

We searched for Yap1-binding sites (TTACTAA or TGACTAA) in the sequences upstream of the target genes we identified (*48*). About two-thirds of the genes that were induced by more than threefold upon Yap1 overexpression had one or more binding sites within 600 bases upstream of the start codon (Table 1), suggesting that they are directly regulated by Yap1. The absence of canonical Yap1-bind-

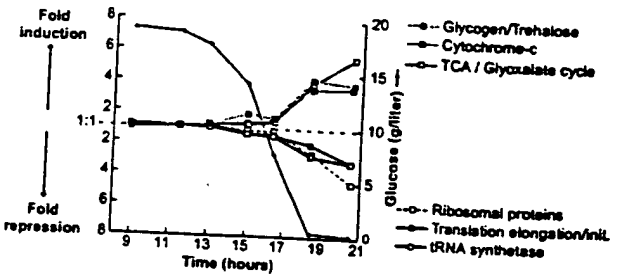

Fig. 4. Coordinated regulation of functionally related genes. The curves represent the average induction or repression ratios for all the genes in each indicated group. The total number of genes in each group was as follows: ribosomal proteins, 112; translation elongation and initiation factors, 25; tRNA synthetases (excluding mitochondial synthetases), 17; glycogen and trehalose synthesis and degradation, 15; cytochrome c oxidase and reductase proteins, 19; and TCA- and glyoxylate-cycle enzymes, 24.

Table 1. Genes induced by *YAP1* overexpression. This list includes all the genes for which mRNA levels increased by more than twofold upon *YAP1* overexpression in both of two duplicate experiments, and for which the average increase in mRNA level in the two experiments was greater than threefold (*50*). Positions of the canonical Yap1 binding sites upstream of the start codon, when present, and the average fold-increase in mRNA levels measured in the two experiments are indicated.

ORF	Distance of Yap1 site from ATG	Gene	Description	Fold-increase
YNL331C			Putative aryl-alcohol reductase	12.9
YKL071W	162–222 (5 sites)		Similarity to bacterial csgA protein	10.4
YML007W		YAP1	Transcriptional activator involved in oxidative stress response	9.8
YFL056C	223, 242		Homology to aryl-alcohol dehydrogenases	9.0
YLL060C	98		Putative glutathione transferase	7.4
YOL165C	266		Putative aryl-alcohol dehydrogenase (NADP+)	7.0
YCR107W			Putative aryl-alcohol reductase	6.5
YML116W	409	ATR1	Aminotriazole and 4-nitroquinoline resistance protein	6.5
YBR008C	142, 167, 364		Homology to benomyl/methotrexate resistance protein	6.1
YCLX08C			Hypothetical protein	6.1
YJR155W			Putative aryl-alcohol dehydrogenase	6.0
YPL171C	148, 212	OYE3	NAPDH dehydrogenase (old yellow enzyme), isoform 3	5.8
YLR460C	167, 317		Homology to hypothetical proteins YCR102c and YNL134c	4.7
YKR076W	178		Homology to hypothetical protein YMR251w	4.5
YHR179W	327	OYE2	NAD(P)H oxidoreductase (old yellow enzyme), isoform 1	4.1
YML131W	507		Similarity to *A. thaliana* zeta-crystallin homolog	3.7
YOL126C		MDH2	Malate dehydrogenase	3.3

ing sites upstream of the others may reflect an ability of Yap1 to bind sites that differ from the canonical binding sites, perhaps in cooperation with other factors, or less likely, may represent an indirect effect of Yap1 overexpression, mediated by one or more intermediary factors. Yap1 sites were found only four times in the corresponding region of an arbitrary set of 30 genes that were not differentially regulated by Yap1.

Use of a DNA microarray to characterize the transcriptional consequences of mutations affecting the activity of regulatory molecules provides a simple and powerful approach to dissection and characterization of regulatory pathways and networks. This strategy also has an important practical application in drug screening. Mutations in specific genes encoding candidate drug targets can serve as surrogates for the ideal chemical inhibitor or modulator of their activity. DNA microarrays can be used to define the resulting signature pattern of alterations in gene expression, and then subsequently used in an assay to screen for compounds that reproduce the desired signature pattern.

DNA microarrays provide a simple and economical way to explore gene expression patterns on a genomic scale. The hurdles to extending this approach to any other organism are minor. The equipment required for fabricating and using DNA microarrays (9) consists of components that were chosen for their modest cost and simplicity. It was feasible for a small group to accomplish the amplification of more than 6000 genes in about 4 months and, once the amplified gene sequences were in hand, only 2 days were required to print a set of 110 microarrays of 6400 elements each. Probe preparation, hybridization, and fluorescent imaging are also simple procedures. Even conceptually simple experiments, as we described here, can yield vast amounts of information. The value of the information from each experiment of this kind will progressively increase as more is learned about the functions of each gene and as additional experiments define the global changes in gene expression in diverse other natural processes and genetic perturbations. Perhaps the greatest challenge now is to develop efficient methods for organizing, distributing, interpreting, and extracting insights from the large volumes of data these experiments will provide.

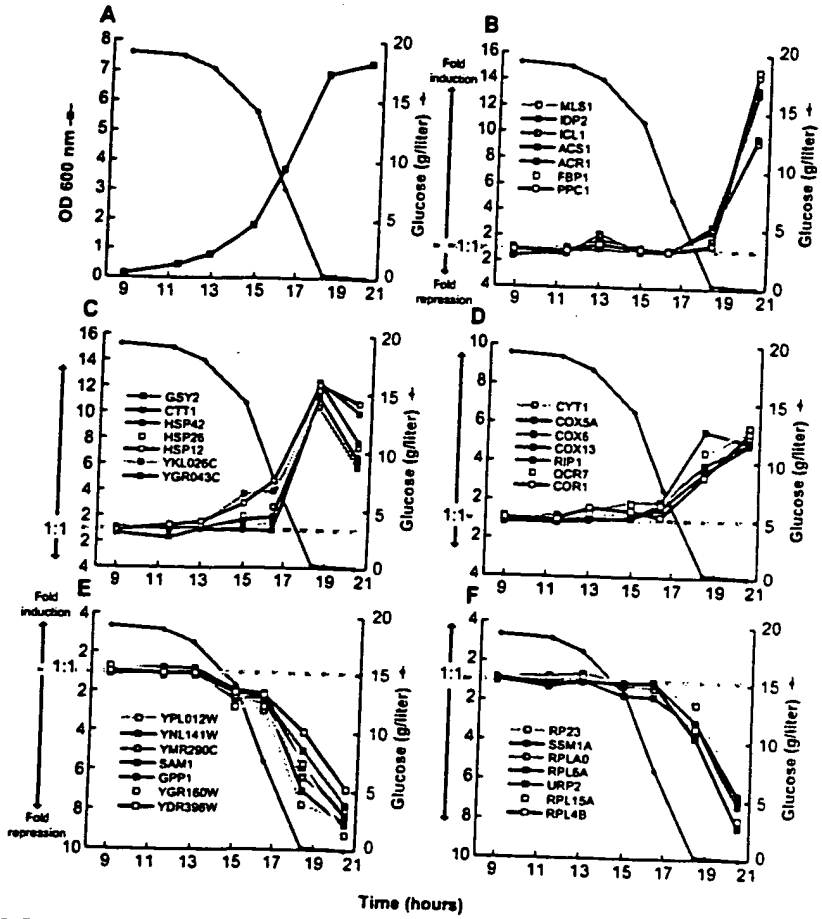

Fig. 5. Distinct temporal patterns of induction or repression help to group genes that share regulatory properties. (A) Temporal profile of the cell density, as measured by OD at 600 nm and glucose concentration in the media. (B) Seven genes exhibited a strong induction (greater than ninefold) only at the last timepoint (20.5 hours). With the exception of *IDP2*, each of these genes has a CSRE UAS. There were no additional genes observed to match this profile. (C) Seven members of a class of genes marked by early induction with a peak in mRNA levels at 18.5 hours. Each of these genes contain STRE motif repeats in their upstream promoter regions. (D) Cytochrome c oxidase and ubiquinol cytochrome c reductase genes. Marked by an induction coincident with the diauxic shift, each of these genes contains a consensus binding motif for the HAP2,3,4 protein complex. At least 17 genes shared a similar expression profile. (E) *SAM1*, *GPP1*, and several genes of unknown function are repressed before the diauxic shift, and continue to be repressed upon entry into stationary phase. (F) Ribosomal protein genes comprise a large class of genes that are repressed upon depletion of glucose. Each of the genes profiled here contains one or more RAP1-binding motifs upstream of its promoter. RAP1 is a transcriptional regulator of most ribosomal proteins.

REFERENCES AND NOTES

1. M. Schena, D. Shalon, R. W. Davis, P. O. Brown, *Science* 270, 467 (1995).
2. D. Shalon, S. J. Smith, P. O. Brown, *Genome Res.* 6, 639 (1996).
3. D. Lashkari, *Proc. Natl. Acad. Sci. U.S.A.*, in press.
4. J. DeRisi *et al.*, *Nature Genet.* 14, 457 (1996).
5. D. J. Lockhart *et al.*, *Nature Biotechnol.* 14, 1675 (1996).
6. M. Chee *et al.*, *Science* 274, 610 (1996).
7. M. Johnston and M. Carlson, in *The Molecular Biology of the Yeast* Saccharomyces: *Gene Expression*, E. W. Jones, J. R. Pringle, J. R. Broach, Eds. (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, 1992), p. 193.
8. Primers for each known or predicted protein coding sequence were supplied by Research Genetics. PCR was performed with the protocol supplied by Research Genetics, using genomic DNA from yeast strain S288C as a template. Each PCR product was verified by agarose gel electrophoresis and was deemed correct if the lane contained a single band of appropriate mobility. Failures were marked as such in the database. The overall success rate for a single-pass amplification of 6116 ORFs was ~94.5%.
9. Glass slides (Gold Seal) were cleaned for 2 hours in a solution of 2 N NaOH and 70% ethanol. After rinsing in distilled water, the slides were then treated with a 1:5 dilution of poly-L-lysine adhesive solution (Sigma) for 1 hour, and then dried for 5 min at 40°C in a vacuum oven. DNA samples from 100-μl PCR reactions were purified by ethanol purification in 96-well microtiter plates. The resulting precipitates were resuspended in 3× standard saline citrate (SSC) and transferred to new plates for arraying. A custom-built arraying robot was used to print on a batch of 110 slides. Details of the design of the microarrayer are available at cmgm.stanford.edu/pbrown. After printing, the microarrays were rehydrated for 30 s in a humid chamber and then snap-dried for 2 s on a hot plate (100°C). The DNA was then ultraviolet (UV)-crosslinked to the surface by subjecting the slides to 60 mJ of energy (Stratagene Stratalinker). The rest of the poly-L-lysine surface was blocked by a 15-min incubation in a solution of 70 mM succinic anhydride dissolved in a solution consisting of 315 ml of 1-methyl-2-pyrrolidinone (Aldrich) and 35 ml of 1 M boric acid (pH 8.0). Directly after the blocking reac-

tion, the bound DNA was denatured by a 2-min incubation in distilled water at −95°C. The slides were then transferred into a bath of 100% ethanol at room temperature, rinsed, and then spun dry in a clinical centrifuge. Slides were stored in a closed box at room temperature until used.

10. YPD medium (8 liters), in a 10-liter fermentation vessel, was inoculated with 2 ml of a fresh overnight culture of yeast strain DBY7286 (MATa, ura3, GAL2). The fermentor was maintained at 30°C with constant agitation and aeration. The glucose content of the media was measured with a UV test kit (Boehringer Mannheim, catalog number 716251). Cell density was measured by OD at 600-nm wavelength. Aliquots of culture were rapidly withdrawn from the fermentation vessel by peristaltic pump, spun down at room temperature, and then flash frozen with liquid nitrogen. Frozen cells were stored at −80°C.
11. Cy3-dUTP or Cy5-dUTP (Amersham) was incorporated during reverse transcription of 1.25 μg of polyadenylated [poly(A)+] RNA, primed by a dT(16) oligomer. This mixture was heated to 70°C for 10 min, and then transferred to ice. A premixed solution, consisting of 200 U Superscript II (Gibco), buffer, deoxyribonucleoside triphosphates, and fluorescent nucleotides, was added to the RNA. Nucleotides were used at these final concentrations: 500 μM for dATP, dCTP, and dGTP and 200 μM for dTTP. Cy3-dUTP and Cy5-dUTP were used at a final concentration of 100 μM. The reaction was then incubated at 42°C for 2 hours. Unincorporated fluorescent nucleotides were removed by first diluting the reaction mixture with of 470 μl of 10 mM tris-HCl (pH 8.0)/1 mM EDTA and then subsequently concentrating the mix to ~5 μl, using Centricon-30 microconcentrators (Amicon).
12. Purified, labeled cDNA was resuspended in 11 μl of 3.5× SSC containing 10 μg poly(dA) and 0.3 μl of 10% SDS. Before hybridization, the solution was boiled for 2 min and then allowed to cool to room temperature. The solution was applied to the microarray under a cover slip, and the slide was placed in a custom hybridization chamber which was subsequently incubated for ~8 to 12 hours in a water bath at 62°C. Before scanning, slides were washed in 2× SSC, 0.2% SDS for 5 min, and then 0.05× SSC for 1 min. Slides were dried before scanning by centrifugation at 500 rpm in a Beckman CS-6R centrifuge.
13. The complete data set is available on the Internet at cmgm.stanford.edu/pbrown/explore/index.html
14. For 95% of all the genes analyzed, the mRNA levels measured in cells harvested at the first and second interval after inoculation differed by a factor of less than 1.5. The correlation coefficient for the comparison between mRNA levels measured for each gene in these two different mRNA samples was 0.98. When duplicate mRNA preparations from the same cell sample were compared in the same way, the correlation coefficient between the expression levels measured for the two samples by comparative hybridization was 0.99.
15. The numbers and identities of known and putative genes, and their homologies to other genes, were gathered from the following public databases: *Saccharomyces* Genome Database (genome-www.stanford.edu), Yeast Protein Database (quest7.proteome.com), and Munich Information Centre for Protein Sequences (speedy.mips.biochem.mpg.de/mips/yeast/index.html).
16. A. Scholer and H. J. Schuller, *Mol. Cell. Biol.* **14**, 3613 (1994).
17. S. Kratzer and H. J. Schuller, *Gene* **161**, 75 (1995).
18. R. J. Haselbeck and H. L. McAlister, *J. Biol. Chem.* **268**, 12116 (1993).
19. M. Fernandez, E. Fernandez, R. Rodicio, *Mol. Gen. Genet.* **242**, 727 (1994).
20. A. Hartig *et al.*, *Nucleic Acids Res.* **20**, 5677 (1992).
21. P. M. Martinez *et al.*, *EMBO J.* **15**, 2227 (1996).
22. J. C. Varela, U. M. Praekelt, P. A. Meacock, R. J. Planta, W. H. Mager, *Mol. Cell. Biol.* **15**, 6232 (1995).
23. H. Ruis and C. Schuller, *Bioessays* **17**, 959 (1995).
24. J. L. Parrou, M. A. Teste, J. Francois, *Microbiology* **143**, 1891 (1997).
25. This expression profile was defined as having an induction of greater than 10-fold at 18.5 hours and less than 11-fold at 20.5 hours.
26. S. L. Forsburg and L. Guarente, *Genes Dev.* **3**, 1166 (1989).
27. J. T. Olesen and L. Guarente, *ibid.* **4**, 1714 (1990).
28. M. Rosenkrantz, C. S. Kell, E. A. Pennell, L. J. Devenish, *Mol. Microbiol.* **13**, 119 (1994).
29. Single-letter abbreviations for the amino acid residues are as follows: A, Ala; C, Cys; D, Asp; E, Glu; F, Phe; G, Gly; H, His; I, Ile; K, Lys; L, Leu; M, Met; N, Asn; P, Pro; Q, Gln; R, Arg; S, Ser; T, Thr; V, Val; W, Trp; and Y, Tyr. The nucleotide codes are as follows: B–C, G, or T; N–G, A, T, or C; R–A or G; and Y–C or T.
30. C. Fondrat and A. Kalogeropoulos, *Comput. Appl. Biosci.* **12**, 363 (1996).
31. D. Shore, *Trends Genet.* **10**, 408 (1994).
32. R. J. Planta and H. A. Raue, *ibid.* **4**, 64 (1988).
33. The degenerate consensus sequence VYCYRNNCMNH was used to search for potential RAP1-binding sites. The exact consensus, as defined by (*30*), is WACAYCCRTACATYW, with up to three differences allowed.
34. S. F. Neuman, S. Bhattacharya, J. R. Broach, *Mol. Cell. Biol.* **15**, 3187 (1995).
35. P. Lesage, X. Yang, M. Carlson, *ibid.* **16**, 1921 (1996).
36. For example, we observed large inductions of the genes coding for *PCK1*, *FBP1* [Z. Yin *et al.*, *Mol. Microbiol.* **20**, 751 (1996)], the central glyoxylate cycle gene *ICL1* [A. Scholer and H. J. Schuller, *Curr. Genet.* **23**, 375 (1993)], and the "aerobic" isoform of acetyl-CoA synthase, *ACS1* [M. A. van den Berg *et al.*, *J. Biol. Chem.* **271**, 28953 (1996)], with concomitant down-regulation of the glycolytic-specific genes *PYK1* and *PFK2* [P. A. Moore *et al.*, *Mol. Cell. Biol.* **11**, 5330 (1991)]. Other genes not directly involved in carbon metabolism but known to be induced upon nutrient limitation include genes encoding cytosolic catalase T *CTT1* [P. H. Bissinger *et al.*, *ibid.* **9**, 1309 (1989)] and several genes encoding small heat-shock proteins, such as *HSP12*, *HSP26*, and *HSP42* [I. Farkas *et al.*, *J. Biol. Chem.* **266**, 15602 (1991); U. M. Praekelt and P. A. Meacock, *Mol. Gen. Genet.* **223**, 97 (1990); D. Wotton *et al.*, *J. Biol. Chem.* **271**, 2717 (1996)].
37. The levels of induction we measured for genes that were expressed at very low levels in the uninduced state (notably, *FBP1* and *PCK1*) were generally lower than those previously reported. This discrepancy was likely due to the conservative background subtraction method we used, which generally resulted in overestimation of very low expression levels (*46*).
38. Cross-hybridization of highly related sequences can also occasionally obscure changes in gene expression, an important concern where members of gene families are functionally specialized and differentially regulated. The major alcohol dehydrogenase genes, *ADH1* and *ADH2*, share 88% nucleotide identity. Reciprocal regulation of these genes is an important feature of the diauxic shift, but was not observed in this experiment, presumably because of cross-hybridization of the fluorescent cDNAs representing these two genes. Nevertheless, we were able to detect differential expression of closely related isoforms of other enzymes, such as *HXK1/HXK2* (77% identical) [P. Herrero *et al.*, *Yeast* **11**, 137 (1995)], *MLS1/DAL7* (73% identical) (*20*), and *PGM1/PGM2* (72% identical) [D. Oh, J. E. Hopper, *Mol. Cell. Biol.* **10**, 1415 (1990)], in accord with previous studies. Use in the microarray of deliberately selected DNA sequences corresponding to the most divergent segments of homologous genes, in lieu of the complete gene sequences, should relieve this problem in many cases.
39. F. E. Williams, U. Varanasi, R. J. Trumbly, *Mol. Cell. Biol.* **11**, 3307 (1991).
40. D. Tzamarias and K. Struhl, *Nature* **369**, 758 (1994).
41. Differences in mRNA levels between the tup1Δ and wild-type strain were measured in two independent experiments. The correlation coefficient between the complete sets of expression ratios measured in these duplicate experiments was 0.83. The concordance between the sets of genes that appeared to be induced was very high between the two experiments. When only the 355 genes that showed at least a twofold increase in mRNA in the tup1Δ strain in either of the duplicate experiments were compared, the correlation coefficient was 0.82.
42. The tup1Δ mutation consists of an insertion of the LEU2 coding sequence, including a stop codon, between the ATG of *TUP1* and an Eco RI site 124 base pairs before the stop codon of the *TUP1* gene.
43. L. R. Kowalski, K. Kondo, M. Inouye, *Mol. Microbiol.* **15**, 341 (1995).
44. M. Viswanathan, G. Muthukumar, Y. S. Cong, J. Lenard, *Gene* **148**, 149 (1994).
45. D. Hirata, K. Yano, T. Miyakawa, *Mol. Gen. Genet.* **242**, 250 (1994).
46. A. Gutierrez, L. Caramelo, A. Prieto, M. J. Martinez, A. T. Martinez, *Appl. Environ. Microbiol.* **60**, 1783 (1994).
47. A. Muheim *et al.*, *Eur. J. Biochem.* **195**, 369 (1991).
48. J. A. Wemmie, M. S. Szczypka, D. J. Thiele, W. S. Moye-Rowley, *J. Biol. Chem.* **269**, 32592 (1994).
49. Microarrays were scanned using a custom-built scanning laser microscope built by S. Smith with software written by N. Ziv. Details concerning scanner design and construction are available at cmgm.stanford.edu/pbrown. Images were scanned at a resolution of 20 μm per pixel. A separate scan, using the appropriate excitation line, was done for each of the two fluorophores used. During the scanning process, the ratio between the signals in the two channels was calculated for several array elements containing total genomic DNA. To normalize the two channels with respect to overall intensity, we then adjusted photomultiplier and laser power settings such that the signal ratio at these elements was as close to 1.0 as possible. The combined images were analyzed with custom-written software. A bounding box, fitted to the size of the DNA spots in each quadrant, was placed over each array element. The average fluorescent intensity was calculated by summing the intensities of each pixel present in a bounding box, and then dividing by the total number of pixels. Local area background was calculated for each array element by determining the average fluorescent intensity for the lower 20% of pixel intensities. Although this method tends to underestimate the background, causing an underestimation of extreme ratios, it produces a very consistent and noise-tolerant approximation. Although the analog-to-digital board used for data collection possesses a wide dynamic range (12 bits), several signals were saturated (greater than the maximum signal intensity allowed) at the chosen settings. Therefore, extreme ratios at bright elements are generally underestimated. A signal was deemed significant if the average intensity after background subtraction was at least 2.5-fold higher than the standard deviation in the background measurements for all elements on the array.
50. In addition to the 17 genes shown in Table 1, three additional genes were induced by an average of more than threefold in the duplicate experiments, but in one of the two experiments, the induction was less than twofold (range 1.6- to 1.9-fold).
51. We thank H. Bennett, P. Spellman, J. Ravetto, M. Eisen, R. Pillai, B. Dunn, T. Ferea, and other members of the Brown lab for their assistance and helpful advice. We also thank S. Friend, D. Botstein, S. Smith, J. Hudson, and D. Dolginow for advice, support, and encouragement; K. Struhl and S. Chatterjee for the Tup1 deletion strain; L. Fernandes for helpful advice on Yap1; and S. Klapholz and the reviewers for many helpful comments on the manuscript. Supported by a grant from the National Human Genome Research Institute (NHGRI) (HG00450), and by the Howard Hughes Medical Institute (HHMI). J.D.R. was supported by the HHMI and the NHGRI. V.R. was supported in part by an Institutional Training Grant in Genome Science (T32 HG00044) from the NHGRI. P.O.B. is an associate investigator of the HHMI.

5 September 1997; accepted 22 September 1997

Drug target validation and identification of secondary drug target effects using DNA microarrays

MATTHEW J. MARTON[1], JOSEPH L. DERISI[2], HOLLY A. BENNETT[1], VISHWANATH R. IYER[2], MICHAEL R. MEYER[1], CHRISTOPHER J. ROBERTS[1], ROLAND STOUGHTON[1], JULJA BURCHARD[1], DAVID SLADE[1], HONGYUE DAI[1], DOUGLAS E. BASSETT, JR[1]., LELAND H. HARTWELL[3], PATRICK O. BROWN[2] & STEPHEN H. FRIEND[1]

[1]*Rosetta Inpharmatics, 12040 115th Avenue NE, Kirkland, Washington 98034, USA*
[2]*Department of Biochemistry, Stanford University School of Medicine, Howard Hughes Medical Institute Stanford, California 94305-5428, USA*
[3]*Fred Hutchinson Cancer Research Center, 1100 Fairview Avenue N., Seattle, Washington 98109, USA*
Correspondence should be addressed to S.H.F; email: sfriend@rosetta.org

We describe here a method for drug target validation and identification of secondary drug target effects based on genome-wide gene expression patterns. The method is demonstrated by several experiments, including treatment of yeast mutant strains defective in calcineurin, immunophilins or other genes with the immunosuppressants cyclosporin A or FK506. Presence or absence of the characteristic drug 'signature' pattern of altered gene expression in drug-treated cells with a mutation in the gene encoding a putative target established whether that target was required to generate the drug signature. Drug dependent effects were seen in 'targetless' cells, showing that FK506 affects additional pathways independent of calcineurin and the immunophilins. The described method permits the direct confirmation of drug targets and recognition of drug-dependent changes in gene expression that are modulated through pathways distinct from the drug's intended target. Such a method may prove useful in improving the efficiency of drug development programs.

Good drugs are potent and specific; that is, they must have strong effects on a specific biological pathway and minimal effects on all other pathways. Confirmation that a compound inhibits the intended target (drug target validation) and the identification of undesirable secondary effects are among the main challenges in developing new drugs. Comprehensive methods that enable researchers to determine which genes or activities are affected by a given drug might improve the efficiency of the drug discovery process by quickly identifying potential protein targets, or by accelerating the identification of compounds likely to be toxic. DNA microarray technology, which permits simultaneous measurement of the expression levels of thousands of genes, provides a comprehensive framework to determine how a compound affects cellular metabolism and regulation on a genomic scale[1-11]. DNA microarrays that contain essentially every open reading frame (ORF) in the *Saccharomyces cerevisiae* genome have already been used successfully to explore the changes in gene expression that accompany large changes in cellular metabolism or cell cycle progression[7-10].

In the modern drug discovery paradigm, which typically begins with the selection of a single molecular target, the ideal inhibitory drug is one that inhibits a single gene product so completely and so specifically that it is as if the gene product were absent. Treating cells with such a drug should induce changes in gene expression very similar to those resulting from deleting the gene encoding the drug's target. Here we have compared the genome-wide effects on gene expression that result from deletions of various genes in the budding yeast *S. cerevisiae* to the effects on gene expression that result from treatment with known inhibitors of those gene products. Using the calcineurin signaling pathway as a model system, we tested an approach that permits identification of genes that encode proteins specifically involved in pathways affected by a drug. The FK506 characteristic pattern, or 'signature', of altered gene expression was not observed in mutant cells lacking proteins inhibited by FK506 (for example, a calcineurin or FK506-binding-protein mutant strain), but was observed in mutants deleted for genes in pathways unrelated to FK506 action (for example, a cyclophilin mutant strain). Conversely, the cyclosporin A (CsA) signature was not observed in CsA-treated calcineurin or cyclophilin mutant strains, but was seen in an FK506-binding-protein mutant strain treated with CsA. The method also demonstrates that FK506, a clinically used immunosuppressant, has 'off-target' effects that are independent of its binding to immunophilins. Thus, the approach we describe may provide a way to identify the pathways altered by a drug and to detect drug effects mediated through unintended targets.

Null mutants phenocopy drug-treated cells on a genomic scale

To test whether a null mutation in a drug target serves as a model of an ideal inhibitory drug, we examined the effects on gene expression associated with pharmacological or genetic inhibition of calcineurin function. Calcineurin is a highly conserved calcium- and calmodulin-activated serine/threonine protein phosphatase implicated in diverse processes dependent on calcium signaling[12-13]. In budding yeast, calcineurin is required for intracellular ion homeostasis[14], for adaptation to prolonged mating pheromone treatment[15] and in the regulation of

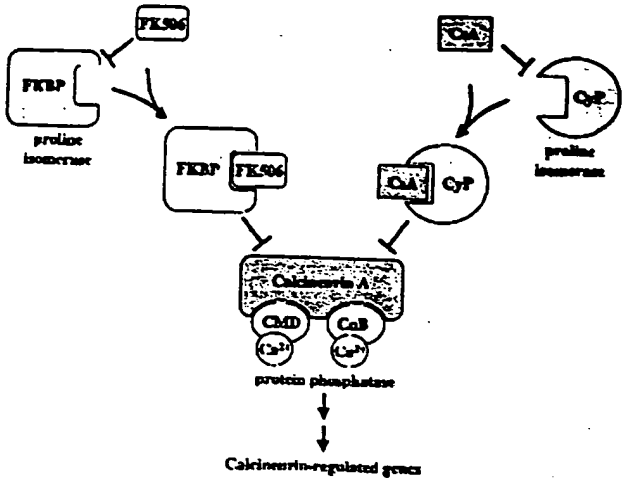

Fig. 1 Model of antagonism of the calcineurin signaling pathway mediated by FK506 and cyclosporin A (CsA). Calcineurin activity is composed of a catalytic subunit (calcineurin A, encoded in yeast by the *CNA1* and *CNA2* genes), and calcium-binding regulatory subunits calmodulin (CMD) and calcineurin B (CnB). After entering cells, FK506 and CsA specifically bind and inhibit the peptidyl-proline isomerase activity of their respective immunophilins, FK506 binding proteins (FKBP) and cyclophilins (CyP). The most abundant immunophilins in yeast (Fpr1 and Cph1) are thought to mediate calcineurin inhibition. Drug–immunophilin complexes bind and inhibit the calcium- and calmodulin-stimulated phosphatase calcineurin. Among the substrates of calcineurin are transcriptional activators that act to modulate gene expression.

the onset of mitosis[16]. In mammals, calcineurin has been implicated in T-cell activation[12], in apoptosis[17], in cardiac hypertrophy[18] and in the transition from short-term to long-term memory[19]. In both organisms, calcineurin activity is inhibited by FK506 and CsA, immunosuppressant drugs whose effects on calcineurin are mediated through families of intracellular receptor proteins called immunophilins[12,20] (Fig. 1). To assess the effects of pharmacologic inhibition of calcineurin, wild-type *S. cerevisiae* was grown to early logarithmic phase in the presence or absence of FK506 or CsA. Isogenic cells, from which the genes encoding the catalytic subunits of calcineurin (*CNA1* and *CNA2*) had been deleted[21] (referred to as the *cna* or calcineurin mutant), were grown in parallel, in the absence of the drug. Fluorescently-labeled cDNA was prepared by reverse transcription of polyA$^+$ RNA in the presence of Cy3- or Cy5-deoxynucleotide triphosphates and then hybridized to a microarray containing more than 6,000 DNA probes representing 97% of the known or predicted ORFs in the yeast genome. Simultaneous hybridization of Cy5-labeled cDNA from mock-treated cells and Cy3-labeled cDNA from cells treated with 1 μg/ml FK506 allowed the effect of drug treatment on mRNA levels of each ORF to be determined (Fig. 2*a* and *b* and data not shown). Similarly, effects of the calcineurin mutations on the mRNA levels of each gene were assessed by simultaneous hybridization of Cy5-labeled cDNA from wild-type cells and Cy3-labeled cDNA from the calcineurin mutant strain(Fig. 2*c*). For each comparison of this kind, reported expression ratios are the average of at least two hybridizations in which the Cy3 and Cy5 fluors were reversed to remove biases that may be introduced by gene-specific differences in incorporation of the two fluors (data not shown).

Treatment with FK506 in these growth conditions resulted in a signature pattern of altered gene expression in which mRNA levels of 36 ORFs changed by more than twofold (http://www.rosetta.org). A very similar pattern of altered gene expression was observed when the calcineurin mutant strain was compared to wild-type cells. Comparison of the changes in mRNA expression of each gene resulting from treatment of wild-type cells with FK506 with mRNA expression changes resulting from deletion of the calcineurin genes showed the considerable similarity of the global transcript alterations in response to the two perturbations (Fig. 2*b–d*). Quantification of this similarity using the correlation coefficient (ρ) showed large correlations between the FK506 treatment signature and the calcineurin deletion signature ($\rho = 0.75 \pm 0.03$), as well as the CsA treatment signature ($\rho = 0.94 \pm 0.02$), but not with a randomly selected deletion mutant strain (deleted for the *YER071C* gene; $\rho = -0.07 \pm 0.04$; Fig. 2*e*). The FK506 treatment signature was also compared with those of more than 40 other deletion mutant strains or drug-treatments thought to affect unrelated pathways, and none had statistically significant correlations. These data establish that genetic disruption of calcineurin function provides a close and specific phenocopy of treatment with FK506 or CsA.

To avoid generalizing from a single example, we also compared the effects of treatment of wild-type cells with 3-aminotriazole (3-AT) with the effects of deletion of the *HIS3* gene. *HIS3* encodes imidazoleglycerol phosphate dehydratase, which catalyzes the seventh step of the histidine biosynthetic pathway in yeast[22]; 3-AT is a competitive inhibitor of this enzyme that triggers a large transcriptional amino-acid starvation response[23]. Microarray analysis of wild-type and isogenic *his3*-deficient strains demonstrated the expected large genome-wide transcriptional responses (involving more than 1,000 ORFs) resulting from treatment with 3-AT (Fig. 3*a*) or from *HIS3* deletion (Fig. 3*c*). Quantitative comparison of the 3-AT treatment signature and the *his3* mutant signature showed a high level of correlation ($\rho = 0.76 \pm 0.02$) that even extended to genes that experienced small changes in expression level (Fig. 3*b*). As a negative control, the correlations between the 3-AT treatment signature or the *his3* mutant signature and the calcineurin mutant strain were not statistically significant ($\rho = 0.09 \pm 0.06$ and -0.01 ± 0.04, respectively). That both the calcineurin/FK506 and the *his3*/3-AT comparisons were highly correlated indicates that in many cases the expression profile resulting from a gene deletion closely resembles the expression profile of wild-type cells treated with an inhibitor of that gene's product.

'Decoder' strategy: Drug target validation with deletion mutants

Because pharmacological inhibition of different targets might give similar or identical expression profiles, simple comparison of drug signatures to mutant signatures is unlikely to unambiguously identify a drug's target. To overcome this limitation, an additional 'decoder' step is used. We first compare the expression profile of wild-type drug-treated cells to the expression profiles from a panel of genetic mutant strains, using a correlation coefficient metric. Mutant strains whose expression profile is similar to that of drug-treated wild-type cells are selected and subjected to drug treatment, generating the drug signature in the mutant strain (that is, the mutant drug signature). If the mutated gene encodes a protein involved in a pathway affected by the drug, we expect the drug signature in mutant cells to be different (or absent, for an ideal drug) from the drug signature seen in wild-type cells.

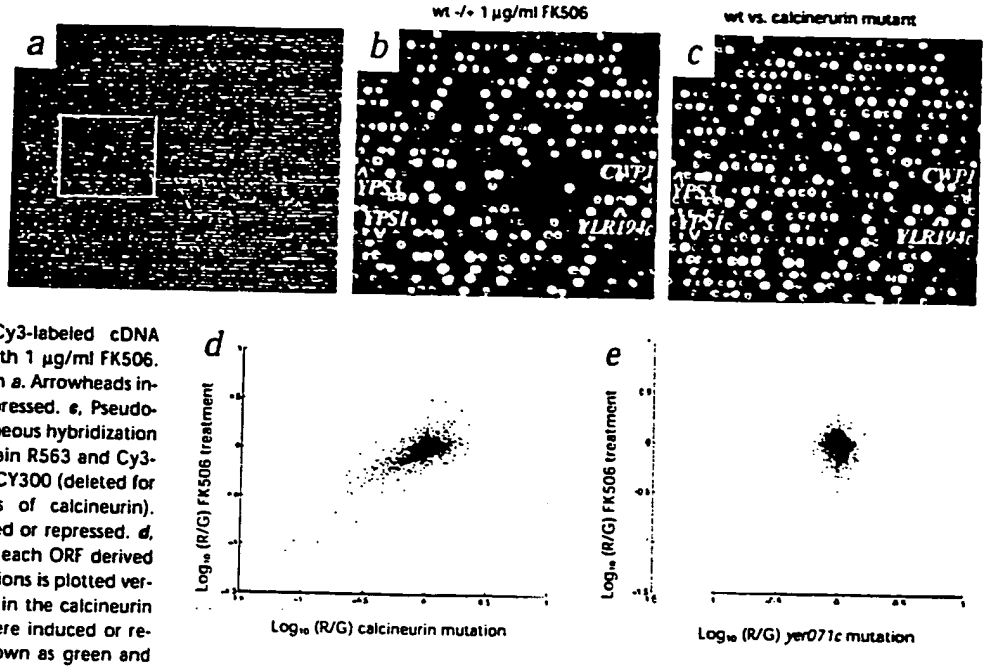

Fig. 2 Expression profiles from FK506-treated wild-type (wt) cells and a calcineurin-disruption mutant strain share a genome-wide correlation. DNA microarray analysis showing changes in gene expression resulting from FK506 treatment (*a* and *b*) or from genetic disruption of genes encoding calcineurin (*c*). *a*, Pseudocolor image of the results of simultaneous hybridization of Cy5-labeled cDNA (red) from mock-treated strain R563 and Cy3-labeled cDNA (green) from strain R563 treated with 1 µg/ml FK506. *b*, Enlarged view of the boxed area in *a*. Arrowheads indicate specific ORFs induced or repressed. *c*, Pseudocolor image of the results of simultaneous hybridization of Cy5-labeled cDNA (red) from strain R563 and Cy3-labeled cDNA (green) from strain MCY300 (deleted for the *CNA1,CNA2* catalytic subunits of calcineurin). Arrows indicate specific ORFs induced or repressed. *d*, The $\log_{10}$ of the expression ratio for each ORF derived from the FK506 treatment hybridizations is plotted versus the $\log_{10}$ of the expression ratio in the calcineurin mutant hybridizations. ORFs that were induced or repressed in both experiments are shown as green and red dots, respectively. *e*, The $\log_{10}$ of the expression ratio for each ORF derived from the FK506 treatment hybridizations is plotted versus the $\log_{10}$ of the expression ratio in the *yer071c* mutant hybridizations. No ORFs were induced or repressed in both experiments.

To illustrate this, we treated the *his3* mutant strain with 3-AT. The signature pattern of altered gene expression resulting from treatment of the mutant strain with 3-AT was much less complex than that of the 3-AT signature in wild-type cells (Fig. 4). This is seen simply by examining plots of mean intensity of the hybridization signal (which approximately reflects level of expression) versus the expression ratio for each ORF (Fig. 4). Genes that were expressed at higher or lower levels in 3-AT treated cells or in *his3* mutant cells are shown as red and green dots, respectively. We analyzed the 3-AT signature in wild-type (Fig. 4a) and *his3* mutant cells (Fig. 4c), as well as the *his3* mutant strain signature (Fig. 4b). Whereas histidine limitation induced by 3-AT induced more than 1,000 transcription-level changes in the wild-type strain, few or no transcript level changes were induced by treatment of the *his3*-deletion strain with 3-AT. This indicates that with the growth conditions used, essentially all of the effects of 3-AT depend on or are mediated through the HIS3 gene product.

Applying this approach to the calcineurin signaling pathway showed the specificity of the method. The calcineurin mutant strain and strains with deletions in the genes encoding the most abundant immunophilins in yeast[12] (*CPH1* and *FPR1*) were treated with either FK506 or CsA to determine the profiles of altered gene expression resulting from drug treatment of the mutant cells (that is, mutant +/- drug). We compared the drug signatures in the mutants to the wild-type drug signature using the correlation coefficient metric (Table 1). Although the signature generated by treatment of wild-type cells with FK506 was highly correlated to the calcineurin mutant strain signature ($\rho = 0.75 \pm 0.03$), it bore no similarity to the profile after treatment of the calcineurin mutant strain with FK506 ($\rho = -0.01 \pm 0.07$). This indicates that FK506 was unable to elicit its normal transcriptional response in the calcineurin mutant strain. Likewise, treatment of the *fpr1* mutant strain with FK506 elicited an expression profile that was not correlated to the FK506 signature in the wild-type strain ($\rho = -0.23 \pm 0.07$), indicating that the *FPR1* gene product is likely to be involved in the pathway affected by FK506. The same was true for the *cna fpr1* mutant strain. In contrast, treatment of the *cph1* mutant strain with FK506 generated an expression profile highly correlated with the wild-type FK506 expression profile ($\rho = 0.79 \pm 0.03$), indicating the *cph1* mutation did not block the mode of action of FK506 and thus is not directly invoived in the pathway affected by FK506. We tabulated the change in expression in response to FK506 in different mutant strains for all ORFs with expression ratios greater than 1.8 in FK506-treated cells or in the calcineurin mutant strain (Fig. 5*a*).The calcineurin mutant strain signature and the FK506 responses in wild-type and the *cph1* mutant strain are similar, and there are no transcript-level changes (seen in black) for treatment of the calcineurin, *fpr1* and *cna fpr1* mutant strains with FK506 (Fig. 5*a*).

Similar experiments and analyses with CsA provided further validation of this approach. The expression profile elicited by treatment of wild-type cells with CsA was highly corre-

Table 1 Signature correlation of expression ratios as a result of FK506 treatment in various mutant strains

	wild-type +/–FK506	*cna* +/–FK506	*fpr1* +/–FK506	*cna fpr1* +/–FK506	*cph1* +/–FK506
wild-type +/- FK506	0.93 ± 0.04	–0.01 ± 0.07	–0.23 ± 0.07	0.12 ± 0.07	0.79 ± 0.03

Signature correlation shows the absence of the FK506 signature specifically in the calcineurin (*cna*) and *fpr1* (major FK506 binding protein) deletion mutants. *cna* represents the mutant with deletions of the catalytic subunits of calcineurin, *CNA1* and *CNA2*. The correlation coefficient reported in the first column represents the correlation between two pairs of hybridizations from independent wild-type +/- FK506 experiments.

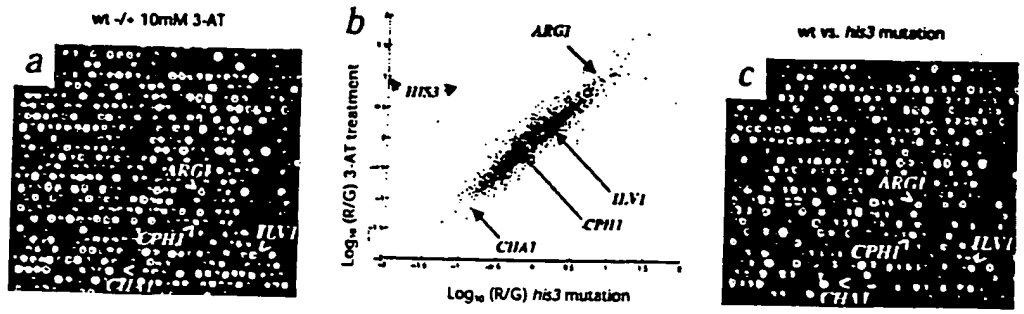

Fig. 3 Expression profiles from a *his3* mutant strain and wild-type (wt) cells treated with 3-AT share a genome-wide correlation. DNA microarray analysis showing changes in gene expression resulting from 3-AT treatment (*a*) or from genetic disruption of the *HIS3* gene (*c*). *a*, Pseudo-color image of the results of simultaneous hybridization of Cy5-labeled cDNA (red) from mock-treated wild-type strain R491 and Cy3-labeled cDNA (green) from strain R491 treated with 10 mM 3-AT. *b*, Plot of the $\log_{10}$ of the expression ratio for each ORF derived from the 3-AT treatment hybridizations is plotted versus the $\log_{10}$ of the expression ratio in the *his3* mutant hybridizations. ORFs that were induced or repressed in both experiments are shown as green and red dots, respectively. The correlation of expression ratios applies not only to genes with large expression ratios (for example, *CHA1* and *ARG1*), but also extends to genes with expression ratios less than 2 (for example, *ILV1* and *CPH1*). *ILV1* is induced 1.9-fold and 1.5-fold, and *CPH1* is downregulated 1.9-fold and 1.7-fold, in cells treated with 3-AT and *his3* mutant cells, respectively. Two ORFs do not fall on the line x = y. The leftmost point is the *HIS3* data point, which is induced by 3-AT treatment but which is not absent from the *his3* mutant strain. The other point is *YOR203w*. Both data points are labeled *HIS3* because hybridization to *YOR203w* is most likely due to *HIS3* mRNA, as *YOR203w* overlaps the *HIS3* open reading frame. *c*, Pseudo-color image of the results of simultaneous hybridization of Cy5-labeled cDNA (red) from wild-type strain R491 and Cy3-labeled cDNA (green) from strain R1226, deleted for the *HIS3* gene. Arrowheads indicate specific ORFs induced or repressed.

lated to the profile elicited by mutation of the calcineurin genes ($\rho = 0.71 \pm 0.04$), but did not correlate with the expression profile resulting from treatment of the calcineurin mutant strain with CsA ($\rho = -0.05 \pm 0.07$; Table 2), indicating that the genetic deletion of calcineurin interfered with the ability of CsA to elicit its normal transcriptional response. Likewise, the CsA signature was essentially absent in CsA-treated *cph1* mutant cells, and the expression profile of CsA-treated *cph1* mutant cells correlated poorly to that of CsA-treated wild-type cells ($\rho = 0.18 \pm 0.07$). Thus, the *CPH1* gene product was required for the CsA response seen in wild-type cells. Conversely, treatment of *fpr1* mutant cells with CsA resulted in an expression pattern very similar to the profile of CsA-treated wild-type cells ($\rho = 0.77 \pm 0.03$), indicating that *FPR1* was not necessary for the CsA-mediated effects. Analysis of individual ORFs affected by CsA and their expression ratios over the entire set of experiments confirmed that *CPH1* and the genes encoding calcineurin, but not *FPR1*, are necessary for the wild-type CsA response (Fig. 5*b*). The observation that the profiles resulting from FK506 or CsA drug treatment are similar to that of the calcineurin deletion mutant strain might allow the prediction that calcineurin was involved in the pathway affected by these drugs. But because the expression profile of the *fpr1* mutant strain did not bear a strong similarity to the wild-type drug expression profile for FK506, it is obvious that the drug treatment of the mutant strains was necessary to identify Fpr1, but not Cph1, as a potential FK506 drug target. In the same way, the 'decoder' strategy was necessary to identify Cph1, but not Fpr1, as a potential drug target for CsA.

'Decoder' approach can identify secondary drug effects

For a drug that has a single biochemical target, the strategy outlined above may be useful in target validation. In many cases, however, a compound may affect multiple pathways and elicit a very complex signature. 'Decoding' such a complex signature

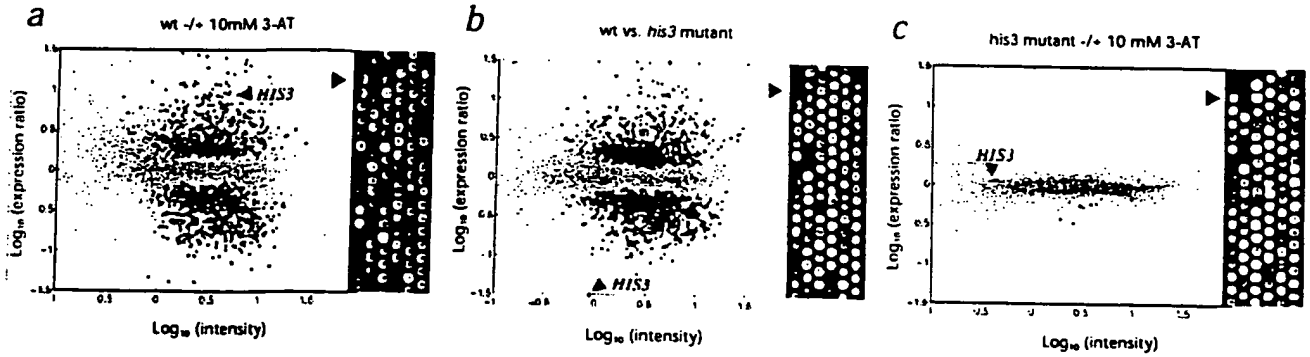

Fig. 4 Treatment of the *his3* mutant strain with 3-AT shows nearly complete loss of 3-AT signature. A plot of the $\log_{10}$ of the mean intensity of hybridization for each ORF versus the $\log_{10}$ of its expression ratio for each experiment is shown next to a pseudo-color image of a representative portion of the microarray. ORFs that are induced or repressed at the 95% confidence level are shown in green and red, respectively. *a*, Expression profile from treatment of the wild-type (wt) strain with 3-AT. Cy5-labeled cDNA (red) from mock-treated strain R491 and Cy3-labeled cDNA (green) from strain R491 treated with 10 mM 3-AT. *b*, Expression profile from the *his3* deletion strain. Cy5-labeled cDNA (red) from strain R491 and Cy3-labeled cDNA (green) from strain R1226, deleted for the *HIS3* gene. *c*, Expression profile of treatment of the *his3* deletion strain with 3-AT. Cy3-labeled cDNA (red) from *his3*-deleted strain R1226 and Cy5-labeled cDNA (green) from strain R1226 treated with 10 mM 3-AT. Arrowheads indicate the DNA probe and data point corresponding to the *HIS3* gene. The blue dashed line represents the threshold below which errors tend to increase rapidly because spot intensities are not sufficiently above background intensity.

Table 2 Signature correlation f expression ratios as a result of CsA treatment in various mutant strains

	wild-type +/–CsA	*cna* +/–CsA	*fpr1* +/–CsA	*cna cph1* +/–CsA	*cph1* +/–CsA
wild-type +/– CsA	0.94 ± 0.04	–0.05 ± .07	0.77 ± 0.03	–0.11 ± 0.07	0.18 ± 0.07

Signature correlation shows the absence of the CsA signature specifically in the calcineurin (*cna*) and *cph1* (cyclophilin) deletion mutants. *cna* represents the mutant with deletions of the catalytic subunits of calcineurin, *CNA1* and *CNA2*. The correlation coefficient reported in the first column represents the correlation between two pairs of hybridizations from independent wild-type +/– CsA experiments.

into the effects mediated through the intended target (the 'on-target signature') and those mediated through unintended targets (the 'off-target' signature) might be useful in evaluating a compound's specificity. Our 'decoder' strategy is based on the premise that 'off-target' signature should be insensitive to the genetic disruption of the primary target.

To determine whether the 'decoder' approach could identify an 'off-target' profile, we looked for a drug-responsive gene whose expression is insensitive to deletion of the primary target. To increase the likelihood of observing such genes, the same strains described in Tables 1 and 2 were treated with higher concentrations (50 μg/ml) of FK506. This led to a much more complex expression profile in wild-type cells, indicating that at this higher concentration, FK506 was inhibiting or activating additional targets. Several of the ORFs in this expanded FK506-induced expression profile were not affected by the calcineurin, *cph1* or *fpr1* mutations, as drug treatment of these mutant strains did not block their presence in the FK506 expression signature (Fig. 6). This indicates that FK506 was triggering changes in transcript levels of many genes through pathways independent of calcineurin, *CPH1* and *FPR1*. Many of the upregulated ORFs in the 'off-target' pathway were genes reported to be regulated by the transcriptional activator Gcn4 (ref. 24). In some strains, a reporter gene under *GCN4* control was induced in response to FK506 treatment[25]. To determine whether *GCN4* is involved in this pathway that is independent of calcineurin, *CPH1* and *FPR1*, we analyzed the effects of treatment with high-dose FK506 on global gene expression in a strain with a *GCN4* deletion (Fig. 6). Of the 41 ORFs with calcineurin-independent expression ratios greater than 4, 32 were not induced in the *gcn4* mutant, indicating that their induction by FK506 was *GCN4*-dependent. Not all *GCN4*-regulated genes were induced by FK506. This FK506-induced subset of *GCN4*-regulated genes may be those most sensitive to subtle changes in Gcn4 levels, or perhaps other regulatory circuits prevent FK506 activation of some *GCN4*-regulated genes. Seven of the remaining nine ORFs induced by FK506 were independent of both the calcineurin and *GCN4* pathways. The simplest explanati n is that FK506 inhibits or activates additional pathways. Members f this class include *SNQ2* and *PDR5*, genes that encode drug efflux pumps with structural homology to mammalian multiple drug resistance proteins[26]. FK506 may interact directly with Pdr5 to inhibit its function[27]. Our results indicate that treatment with FK506 leads to four-fold-to-sixfold induction of *PDR5* mRNA levels. *YOR1*, another gene that can confer drug resistance, is also induced threefold-to-fourfold by FK506. Thus, drug treatment of strains with mutations in the primary targets can prove useful in identifying effects mediated by secondary drug targets, including the nature and extent of newly discovered and previously unsuspected pathways affected by the drug.

We describe here a method for drug target validation and the identification of secondary drug target effects that uses DNA microarrays to survey the effects of drugs on global gene expression patterns. We established that genetic and pharmacologic inhibition of gene function can result in extremely similar changes in gene expression. We also demonstrated that one can confirm a potential drug target by treating a deletion mutant defective in the gene encoding the putative target. Drug-mediated signatures from strains with mutations in pathways or processes directly or indirectly affected by the drug bore little or

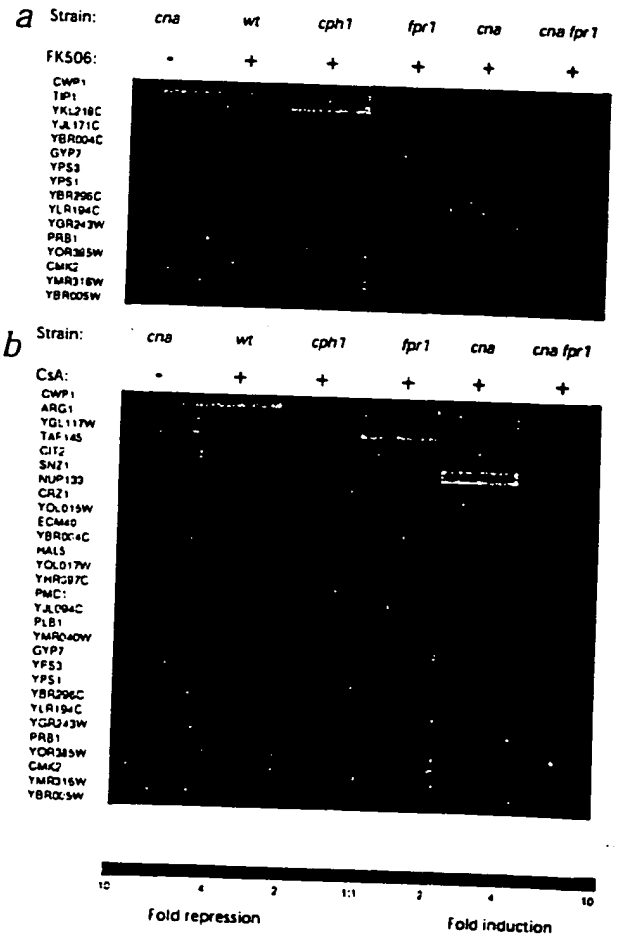

Fig. 5 Response of FK506 and CsA signature genes in strains with deletions in different genes. Genes with expression ratios greater than a factor of 1.8 in response to treatment with 1 μg/ml FK506 (*a*) or 50 μg/ml CsA (*b*) are listed (left side) and their expression ratios in the indicated strain are shown on the green (induction)–red (repression) color scale. *a*, Calcineurin (*cna*) mutant and FK506 treatment signature genes are in the first two columns. Almost all FK506 signature genes have expression ratios near unity in deletion strains involved in pathways affected by FK506 (calcineurin, *fpr1* and *cna fpr1* mutants) but not in deletion strains in unrelated pathways (*cph1*). *b*, Calcineurin (*cna*) mutant and CsA treatment signature genes are in the first two columns. Almost all CsA signature genes have expression ratios near unity in deletion strains involved in pathways affected by CsA (calcineurin, *cph1* and *cna cph1* mutants) but not in deletion strains in unrelated pathways (*fpr1*).

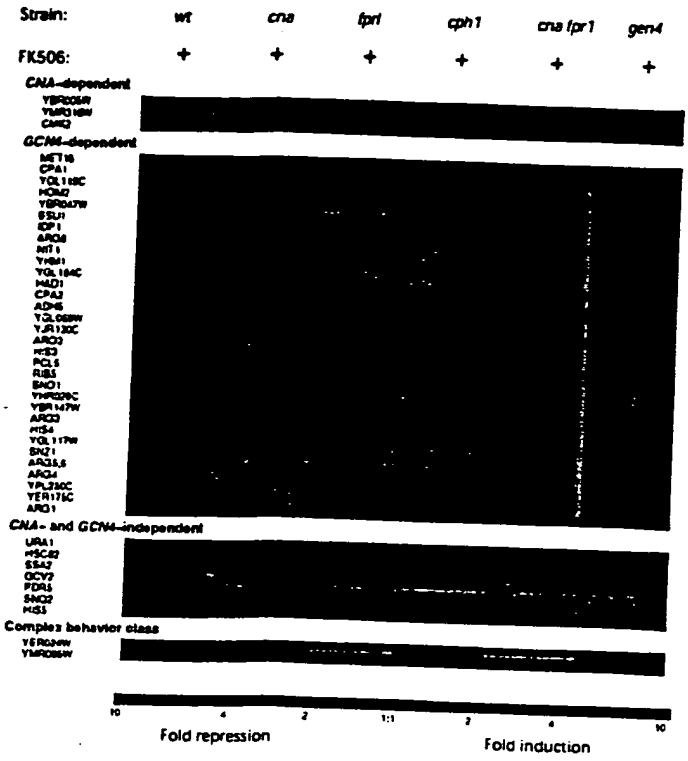

Fig. 6 Response of FK506 signature genes in strains with deletions in different genes. Genes with expression ratios greater than a factor of 4 in at least one experiment are listed and their expression ratios in the indicated strain are shown in the green (induction)–red (repression) color scale. The genes have been divided into classes corresponding to these expected behaviors: '*CNA*-dependent' genes respond to FK506 (50 μg/ml) except when either calcineurin genes or *FPR1* or both are deleted; '*GCN4*-dependent' genes respond to FK506 except when *GCN4* is deleted. These genes still respond to FK506 when calcineurin genes or *FPR1* or *CPH1* are deleted; that is, their responses are not mediated by calcineurin, Cph1, or Fpr1. '*CNA*- and *GCN4*-independent' genes respond to FK506 in all deletion strains tested. A 'complex behavior' class is provided for those genes that did not match the model of FK506 response mediated through calcineurin or Fpr1 or separately through Gcn4.

no similarity to the wild-type drug expression profile. In contrast, drug-mediated signatures from strains with mutations in genes involved in pathways unrelated to the drug's action showed extensive similarity to the wild-type drug signature. By applying this approach to a drug that affects multiple pathways (FK506), we were able to decode a complex signature into component parts, including the identification of an 'off-target' signature that was mediated through pathways independent of calcineurin or the Fpr1 immunophilin.

Discussion

It is well-established that high-throughput biochemical screening can identify potent inhibitory compounds against a given target. The 'decoder' approach described here complements this process by evaluating the equally important property of specificity: the tendency of a compound to inhibit pathways other than that of its intended target. The ability to observe such 'off-target' effects will likely be useful in several ways. Profiling compounds with known toxicities will allow the development of a database of expression changes associated with particular toxicities. Recognition of potential toxicities in the 'off-target' signatures of otherwise promising compounds then may allow earlier identification of those likely to fail in clinical trials. Comparing the extent and peculiarities of 'off-target' signatures of promising drug candiates could provide a new way to group compounds by their effects on secondary pathways, even before those effects are understood. This may prove to be an alternative, potentially more effective, way to select compounds for animal and clinical trials. Some drugs are more effective against a related protein than against the originally intended target. Sildenafil (Viagra™), for example, was initially developed as a phosphodiesterase inhibitor to control cardiac contractility, but was found to be highly specific for phosphodiesterase 5, an isozyme whose inhibition overcomes defects in penile erection. It is possible that application of the 'decoder' to other compounds may show that they too have a potent activity against a target distinct from their intended target.

The ability to decode drug effects is dependent on the availability of functionally 'targetless' cells. In yeast, this is being achieved by systematically disrupting each yeast gene (*Saccharomyces* Deletion Consortium: http://sequence-www.stanford.edu/group/yeast_deletion_project/deletion.html). Efforts are underway to obtain expression profiles from each deletion mutant strain. Determining signatures resulting from inactivation of essential genes presents a unique problem, but it may be possible to do so by examining heterozygotes or by using a controllable promoter to reduce expression of the essential gene. Although it is already feasible to test several compounds in dozens of yeast strains, another challenge for the 'decoder' strategy will be the efficient selection of the mutants with deletions in genes most likely to encode the intended drug target. The signature correlation plots described are one metric that could be used as part of that selection process, but others need to be explored. Applying the 'decoder' to mammalian cells presents additional challenges. It is considerably more difficult to isolate functionally 'targetless' cells. Strategies involving titratable promoters, known specific inhibitors, anti-sense RNAs, ribozymes, and methods of targeting specific proteins for degradation are possible and should be tested. Another limitation is that not all cell types express the same set of genes and therefore 'off-target' effects may be different in different cell types. In addition, applying the 'decoder' to human cells will also require technical improvements that allow expression profiling from a small number of cells. Even the broader question of whether the insensitivity of 'off-target' signatures to the disruption of the main target is the exception or the rule can only be answered by the accumulation of more data. Barkai and Leibler, however, have argued in favor of robustness of biological networks, indicating that drug perturbations ('off-target' signatures) may be robust even when the system is subjected to another perturbation (such as a genetic disruption)(ref. 28). Many practical developments will be necessary if the 'decoder' concept is to be broadly applied.

Expression arrays have been used mainly as an initial screen for genes induced in a particular tissue or process of interest by focusing on genes with large expression ratios. We have found, however, that effort to refine experimental protocols and repeat experiments increases the reliability of the data and permits new applications. For example, it provides a larger set

Table 3 Yeast strains used

Strain	Relevant genotype	Reference
YPH499	*Mata ura3-52 lys2-801 ade2-101 trp1-Δ63 his3-Δ200 leu2-Δ1*	(34)
R563	*Mata ura3-52 lys2-801 ade2-101 trp1-Δ63 his3-Δ200 leu2-Δ1 his3::HIS3*	(this study)
R558	*Mata ura3-52 lys2-801 ade2-101 trp1-Δ63 his3-Δ200 leu2-Δ1 fpr1::HIS3*	(this study)
R567	*Mata ura3-52 lys2-801 ade2-101 trp1-Δ63 his3-Δ200 leu2-Δ1 cph1::HIS3*	(this study)
MCY300	*Mata ura3-52 lys2-801 ade2-101 trp1-Δ63 his3-Δ200 leu2-Δ1 cna1Δ1::hisG cna2Δ1::HIS3*	(21)
R132	*Mata ura3-52 lys2-801 ade2-101 trp1-Δ63 his3-Δ200 leu2-Δ1 cna1Δ1::hisG cna2Δ1::HIS3 cph1::kanr*	(this study)
R133	*Mata ura3-52 lys2-801 ade2-101 trp1-Δ63 his3-Δ200 leu2-Δ1 cna1Δ1::hisG cna2Δ1::HIS3 fpr1::kanr*	(this study)
R559	*Mata ura3-52 lys2-801 ade2-101 trp1-Δ63 his3-Δ200 leu2-Δ1 his3::HIS3 gcn4::LEU2*	(this study)
BY4719	*Mata trp1-Δ63 ura3-Δ0*	(35)
BY4738	*Matα trp1-Δ 63 ura3-Δ0*	(35)
R491	*Mata/α* BY4719 *X* BY4738	(this study)
BY4728	*Mata his3-Δ200 trp1-Δ63 ura3-Δ0*	(35)
BY4729	*Matα his3-Δ 200 trp1-Δ63 ura3-Δ0*	(35)
R1226	*Mata/α* BY4728 *X* BY4729	(this study)

of genes at higher confidence levels that serve as a more unique signature for a given protein perturbation. In addition, it allows subtle signatures to be detected, when, for example, a protein is only partially inhibited. This may enable clinical monitoring of small changes in protein function in disease or toxicity states before they could otherwise be detected. Because the functions of many genes detected on transcript arrays are known, these microarrays are powerful tools that provide detailed information about a cell's physiology. For example, changes in the flux through a metabolic pathway are reflected in transcriptional changes in genes in the pathway[7]. Furthermore, it may be possible to indirectly measure protein activity levels from expression profiling data (S.F., *et al.*, unpublished data). Thus, although the eventual development of genomic methods allowing the direct measurement of all cellular protein levels will be an important achievement, transcript array technology offers an immediate and robust means of evaluating the effects of various treatments on gene expression and protein function.

Methods

Construction, growth and drug treatment of yeast strains. The strains used in this study (Table 3) were constructed by standard techniques[29]. To construct strain R559, strain R563 was transformed to Leu$^+$ with plasmid pM12 digested by *Sal*I and *Mlu*I (provided by A. Hinnebusch and T. Dever). Strains R132 and R133 were constructed by transforming the bacterial kanamycin resistance cassette[30] flanked by genomic DNA from the *CPH1* and *FPR1* loci, respectively, and selecting for G418-resistant colonies. For experiments with FK506, cells were grown for three generations to a density of 1 x 10^7 cells/ml in YAPD medium (YPD plus 0.004% adenine) supplemented with 10 mM calcium chloride as described[31]. Where indicated, FK506 was added to a final concentration of 1 μg/ml 0.5 h after inoculation of the culture or to 50 μg/ml 1 h before cells were collected. CsA was used at a final concentration of 50 μg/ml. Cells were broken by standard procedures[32] with the following modifications: Cell pellets were resuspended in breaking buffer (0.2 M Tris HCl pH 7.6, 0.5 M NaCl, 10 mM EDTA, 1% SDS), vortexed for 2 min on a VWR multi-tube vortexer at setting 8 in the presence of 60% glass beads (425–600 μm mesh; Sigma) and phenol:chloroform (50:50, volume/volume). After separation of the phases, the aqueous phase was re-extracted and ethanol-precipitated. Poly A$^+$ RNA was isolated by two sequential chromatographic purifications over oligo dT cellulose (New England Biolabs, Beverly, Massachusetts) using established protocols[32].

For experiments using 3-AT, wild-type or *his3/his3* cells were grown to early logarithmic phase in SC medium, pelleted and resuspended in SC medium lacking histidine for 1 hr in the presence or absence of 10 mM 3-AT, as indicated. Cells were harvested and mRNA isolated as above. FK506 was obtained from the Swedish Hospital Pharmacy (Seattle, Washington) and purified to homogeneity by ethyl acetate extraction by J. Simon (Fred Hutchinson Cancer Research Center, Seattle, Washington). CsA was obtained from Alexis Biochemicals (San Diego, California); 3-AT was from Sigma.

Preparation and hybridization of the labeled sample. Fluorescently-labeled cDNA was prepared, purified and hybridized essentially as described[7]. Cy3- or Cy5-dUTP (Amersham) was incorporated into cDNA during reverse transcription (Superscript II; Life Technologies) and purified by concentrating to less than 10 μl using Microcon-30 microconcentrators (Amicon, Houston, Texas). Paired cDNAs were resuspended in 20–26 μl hybridization solution (3 x SSC, 0.75 μg/ml polyA DNA, 0.2% SDS) and applied to the microarray under a 22- x 30-mm coverslip for 6 h at 63 °C, all according to a published method[7].

Fabrication and scanning of microarrays. PCR products containing common 5' and 3' sequences (Research Genetics, Huntsville, Alabama) were used as templates with amino-modified forward primer and unmodified reverse primers to PCR amplify 6,065 ORFs from the *S. cerevisiae* genome. Our first-pass success rate was 94%. Amplification reactions that gave products of unexpected sizes were excluded from subsequent analysis. ORFs that could not be amplified from purchased templates were amplified from genomic DNA. DNA samples from 100-μl reactions were isopropanol-precipitated, resuspended in water, brought to a final concentration of 3x SSC in a total volume of 15 μl, and transferred to 384-well microtiter plates (Genetix Limited, Christchurch, Dorset, England). PCR products were spotted onto 1 x 3-inch polylysine-treated glass slides by a robot built essentially according to defined specifications[2,4,7] (http://cmgm.stanford.edu/pbrown/MGuide). After being printed, slides were processed according to published protocols[7].

Microarrays were imaged on a prototype multi-frame CCD camera in development at Applied Precision (Issaquah, Washington). Each CCD image frame was approximately 2-mm square. Exposure times of 2 s in the Cy5 channel (white light through Chroma 618–648 nm excitation filter, Chroma 657–727 nm emission filter) and 1 s in the Cy3 channel (Chroma 535–560 nm excitation filter, Chroma 570–620 nm emission filter) were done consecutively in each frame before moving to the next, spatially contiguous frame. Color isolation between the Cy3 and Cy5 channels was about 100:1 or better. Frames were 'knitted' together in software to make the complete images. The intensity of spots (about 100 μm) were quantified from the 10-μm pixels by frame-by-frame background subtraction and intensity averaging in each channel. Dynamic range of the resulting spot intensities was typically a ratio of 1,000 between the brightest spots and the background-subtracted additive error level. Normalization between the channels was accomplished by normalizing each channel to the mean intensities of all genes. This procedure is nearly equivalent to normalization between channels using the intensity

ratio of genomic DNA spots[7], but is possibly more robust, as it is based on the intensities of several thousand spots distributed over the array.

Signature correlation coefficients and their confidence limits. Correlation coefficients between the signature ORFs of various experiments were calculated using:

$$\rho = \sum_k x_k y_k / (\sum_k x_k^2 \sum_k y_k^2)^{1/2}$$

where x_k is the $\log_{10}$ of the expression ratio for the k^{th} gene in the x signature, and y_k is the $\log_{10}$ of the expression ratio for the k^{th} gene in the y signature. The summation is over those genes that were either up- or down-regulated in either experiment at the 95% confidence level. These genes each had a less than 5% chance of being actually unregulated (having expression ratios departing from unity due to measurement errors alone). This confidence level was assigned based on an error model which assigns a lognormal probability distribution to each gene's expression ratio with characteristic width based on the observed scatter in its repeated measurements (repeated arrays at the same nominal experimental conditions) and on the individual array hybridization quality. This latter dependence was derived from control experiments in which both Cy3 and Cy5 samples were derived from the same RNA sample. For large numbers of repeated measurements the error reduces to the observed scatter. For a single measurement the error is based on the array quality and the spot intensity.

Random measurement errors in the x and y signatures tend to bias the correlation towards zero. In most experiments, most genes are not significantly affected but do show small random measurement errors. Selecting only the '95% confidence' genes for the correlation calculation, rather than the entire genome, reduces this bias and makes the actual biological correlations more apparent.

Correlations between a profile and itself are unity by definition. Error limits on the correlation are 95% confidence limits based on the individual measurement error bars, and assuming uncorrelated errors[33]. They do not include the bias mentioned above; thus, a departure of ρ from unity does not necessarily mean that the underlying biological correlation is imperfect. However, a correlation of 0.7 ± 0.1, for example, is very significantly different from zero. Small (magnitude of $\rho < 0.2$) but formally significant correlation in the tables and text probably are due to small systematic biases in the Cy5/Cy3 ratios that violate the assumption of independent measurement errors used to generate the 95% confidence limits. Therefore, these small correlation values should be treated as not significant. A likely source of uncorrected systematic bias is the partially corrected scanner detector nonlinearity that differently affects the Cy3 and Cy5 detection channels.

The 1 μg/ml FK506 treatment signature was compared with more than 40 unrelated deletion mutant strain or drug signatures. These control profiles had correlation coefficients with the FK506 profile that were distributed around zero (mean $\rho = -0.03$) with a standard deviation of 0.16 (data not shown), and none had correlations greater than $\rho = 0.38$. Similarly, the calcineurin mutant strain signature correlated well with the CsA treatment signature ($\rho = 0.71 \pm 0.04$) but not with the signatures from the negative controls (mean $\rho = -0.02$ with a standard deviation of 0.18).

Quality controls. End-to-end checks on expression ratio measurement accuracy were provided by analyzing the variance in repeated hybridizations using the same mRNA labeled with both Cy3 and Cy5, and also using Cy3 and Cy5 mRNA samples isolated from independent cultures of the same nominal strain and conditions. Biases undetected with this procedure, such as gene-specific biases presumably due to differential incorporation of Cy3- and Cy5-dUTP into cDNA, were minimized by doing hybridizations in fluor-reversed pairs, in which the Cy3/Cy5 labeling of the biological conditions was reversed in one experiment with respect to the other. The expression ratio for each gene is then the ratio of ratios between the two experiments in the pair. Other biases are removed by algorithmic numerical de-trending. The magnitude of these biases in the absence of de-trending and fluor reversal is typically about 30% in the ratio, but may be as high as twofold for some ORFs.

Expression ratios are based on mean intensities over each spot. Some smaller spots have fewer image pixels in the average. This does not degrade accuracy noticeably until the number of pixels falls below ten, in which case the spot is rejected from the data set. 'Wander' of spot positions with respect to the nominal grid is adaptively tracked in array subregions by the image processing software. Unequal spot 'wander' within a subregion greater than half-a-spot spacing is a difficulty for the automated quantitating algorithms; in this case, the spot is rejected from analysis based on human inspection of the 'wander'. Any spots partially overlapping are excluded from the data set. Less than 1% of spots typically are rejected for these reasons.

Acknowledgments

The authors thank all the members of Rosetta for their contributions to this work. We thank P. Linsley, D. Shoemaker and A. Murray for critical reading of the manuscript, and M. Cyert for providing yeast strains. Work done at Stanford was supported in part by the Howard Hughes Medical Institute, and by a grant to P.O.B from the NHGRI. P.O.B is an assistant investigator of the Howard Hughes Medical Institute.

RECEIVED 13 AUGUST; ACCEPTED 2 OCTOBER 1998

1. Schena, M., Shalon, D., Davis, R.W. & Brown, P.O. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* 270, 467–470 (1995).
2. Schena, M. *et al.* Parallel human genome analysis: microarray-based expression monitoring of 1000 genes. *Proc. Natl. Acad. Sci. USA* 93, 10614–10619 (1996).
3. Shalon, D., Smith, S.J. & Brown, P.O. A DNA microarray system for analyzing complex DNA samples using two-color fluorescent probe hybridization. *Genome Res.* 6, 639–645 (1996).
4. Lockhart, D.J. *et al.* Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nature Biotechnol.* 14, 1675–1680 (1996).
5. DeRisi, J. *et al.* Use of a cDNA microarray to analyse gene expression patterns in human cancer. *Nature Genet.* 14, 457–460 (1996).
6. Heller, R.A. *et al.* Discovery and analysis of inflammatory disease-related genes using cDNA microarrays. *Proc. Natl. Acad. Sci. USA* 94, 2150–2155 (1997).
7. DeRisi, J.L., Iyer, V.R. & Brown, P.O. Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science* 278, 680–686 (1997).
8. Lashkari, D.A. *et al.* Yeast microarrays for genome wide parallel genetic and gene expression analysis. *Proc. Natl. Acad. Sci. USA* 94, 13057–13062 (1997).
9. Wodicka, L., Dong H., Mittman, M. Ho, M.-H. & Lockhart, D.J.. Genome-wide expression monitoring in Saccharomyces cerevisiae. *Nature Biotechnol.* 15, 1359–1367 (1997).
10. Cho, R.J. *et al.* A genome-wide transcriptional analysis of the mitotic cell cycle. *Mol. Cell* 2, 65–73 (1998).
11. Gray, N.S. *et al.* Exploiting chemical libraries, structure, and genomics in the search for kinase inhibitors. *Science* 281, 533–538 (1998).
12. Cardenas, M.E., Lorenz, M., Hemenway, C. & Heitman, J. Yeast as model T cells. *Perspect. Drug Discovery Design* 2, 103–126 (1994).
13. Klee, C.B., Ren, H. & Wang, X. Regulation of the calmodulin-stimulated protein phosphatase, calcineurin. *J. Biol. Chem.* 273, 13367–13370 (1998).
14. Tanida, I., Hasegawa, A., Iida, H., Ohya, Y. & Anraku, Y. Cooperation of calcineurin and vacuolar H(+)-ATPase in intracellular Ca2+ homeostasis of yeast cells. *J. Biol. Chem.* 270, 10113–10119 (1995).
15. Moser, M.J., Geiser, J.R. & Davis, T.N. Ca2+-calmodulin promotes survival of pheromone-induced growth arrest by activation of calcineurin and Ca2+-calmodulin-dependent protein kinase. *Mol. Cell. Biol.* 16, 4824–4831 (1996).
16. Mizunuma, M., Hirata, D., Miyahara, K., Tsuchiya, E. & Miyakawa, T. Role of calcineurin and Mpk1 in regulating the onset of mitosis in budding yeast. *Nature* 392, 303–306 (1998).
17. Yazdanbakhsh, K., Choi, J.W., Li, Y., Lau, L.F. & Choi, Y. Cyclosporin A blocks apoptosis by inhibiting the DNA binding activity of the transcription factor Nur77. *Proc. Natl. Acad. Sci. USA* 92, 437–441 (1995).
18. Molkentin, J.D. *et al.* A calcineurin-dependent transcriptional pathway for cardiac hypertrophy. *Cell* 93, 215–228 (1998)
19. Mansuy, I.M., Mayford, M., Jacob, B., Kandel, E.R. & Bach, M.E. Restricted and regulated overexpression reveals calcineurin as a key component in the transition from short-term to long-term memory. *Cell* 92, 39–49 (1998).
20. Schreiber, S.L. & Crabtree, G.R. The mechanism of action of cyclosporin A and FK506. *Immunol. Today* 13, 136–142 (1992).
21. Cyert, M.S., Kunisawa, R., Kaim, D. & Thorner, J. Yeast has homologs (CNA1 and CNA2 gene products) of mammalian calcineurin, a calmodulin-regulated phosphoprotein phosphatase. *Proc. Natl. Acad. Sci. USA* 88, 7376–7380 (1991).
22. Jones, E.W. & Fink, G.R. in *The Molecular Biology of the Yeast Saccharomyces: Metabolism and Gene Expression* (eds. Strathern, J.N., Jones, E.W. & Broach, J.R.) 181–299 (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York, 1982).
23. Hinnebusch, A. Translational regulation of yeast GCN4. *J. Biol. Chem.* 272, 21661–21664 (1997).
24. Hinnebusch, A.G. in *The Molecular and Cellular Biology of the Yeast*

Saccharomyces: Gene Expression. (eds. Jones, E.W., Pringle, J.R. & Broach, J.R.) 319–414 (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York, 1992).
25. Heitman, J. *et al.* The immunosuppressant FK506 inhibits amino acid import in Saccharomyces cerevisiae. *Mol. Cell. Biol.* 13, 5010–5019 (1993).
26. Balzi, E. & Goffeau, A. Yeast multidrug resistance: the PDR network. *J. Bioenerg. Biomembr.* 27, 71–76 (1995).
27. Egner, R., Rosenthal, F.E., Kralli, A., Sanglard, D. & Kuchler, K. Genetic separation of FK506 susceptibility and drug transport in the yeast Pdr5 ATP-binding cassette multidrug resistance transporter *Mol. Biol. Cell* 9, 523–543 (1998).
28. Barkai, N. & Leiber, S. Robustness in simple biochemical networks. *Nature* 387, 913–917 (1997).
29. Schiestl, R.H., Manivasakam, P. Woods, R.A. & Gietz, R.D. Introducing DNA into yeast by transformation. *Methods: A companion to Methods in Enzymology* 5, 79–85 (1993).
30. Wach, A., Brachat, A., Pohlmann, R. & Philippsen, P. New heterologous modules for classical or PCR-based gene disruptions in Saccharomyces cerevisiae. *Yeast* 10, 1793–1808 (1994).
31. Garrett-Engele, P., Moilanen, B. & Cyert, M.S. Calcineurin, the Ca2+/calmodulin-dependent protein phosphatase, is essential in yeast mutants with cell integrity defects and in mutants that lack a functional vacuolar H(+)-ATPase. *Mol. Cell. Biol.* 15, 4103–4114 (1995).
32. Ausubel, F.M. *et al.* in *Current Protocols in Molecular Biology* 13.12.1–13.12.5 (eds. Ausubel, F.M., *et al.*)(John Wiley & Sons, New York, 1993).
33. Bulmer, M.G. in *Principles of Statistics* 224–225 (Dover Publications, New York, 1979).
34. Sikorski, R.S. & Hieter, P. A system of shuttle vectors and yeast host strains designated for efficient manipulation of DNA in *Saccharomyces cerevisiae*. *Genetics* 122, 19–27 (1989).
35. Brachmann, C.B. *et al.* Designer deletion strains derived from Saccharomyces cerevisiae S288C: A useful set of strains and plasmids for PCR-mediated gene disruption and other applications. *Yeast* 14, 115–132 (1998).

co mosaic viral RNA was obtained by phenol and chloroform extractions of the virus and precipitated from ethanol. CA-NC assembly reactions in the presence of noncognate RNAs were identical to those given in (9). In the absence of RNA, CA-NC cones formed under the following conditions: 300 μM CA-NC, 1 M NaCl, and 50 mM tris-HCl (pH 8.0) at 37°C for 60 min. In the absence of exogenous RNA, neither cones nor cylinders formed at concentrations of 0.5 M NaCl or below. Absorption spectra demonstrated that our CA-NC preparations were not contaminated with *Escherichia coli* RNA (estimated lower detection limit was ~1 base/protein molecule). To control for even lower levels of RNA contamination, we preincubated the CA-NC protein with 0.5 mg/ml ribonuclease A (Type 1-AS, 54 Kunitz U/mg, Sigma) for 1 hour at 4°C, which then formed cones normally.

13. V. Y. Klishko, data not shown.
14. M. Ge and K. Sattler, *Chem. Phys. Lett.* 220, 192 (1994).
15. A. Krishnan *et al.*, *Nature* 388, 451 (1997).
16. L. B. Kong *et al.*, *J. Virol.* 72, 4403 (1998).
17. Assembly mixtures were deposited on holey carbon grids, blotted briefly with filter paper, plunged into liquid ethane, and transferred to liquid nitrogen. Frozen grids were transferred to a Philips 420 TEM equipped with a Gatan cold stage system, and images of particles in vitreous ice were recorded under low dose conditions at 36,000× magnification and ~1.6-μm defocus.
18. J. T. Finch, data not shown.
19. R. A. Crowther, *Proceedings of the Third John Innes Symposium* (1976), pp. 15–25; E. Kellenberger, M. Häner, M. Wurtz, *Ultramicroscopy* 9, 139 (1982); J. Seymore and D. J. DeRosier, *J. Microsc.* 148, 195 (1987).
20. M. V. Nermut, C. Grief, S. Hashmi, D. J. Hockley, *AIDS Res. Hum. Retroviruses* 9, 929 (1993); M. V. Nermut *et al.*, *Virology* 198, 288 (1994); E. Barklis, J. McDermott, S. Wilkens, S. Fuller, D. Thompson, *J. Biol. Chem.* 273, 7177 (1998); E. Barklis *et al.*, *EMBO J.* 16, 1199 (1997); M. Yeager, E. M. Wilson-Kubalek, S. G. Weiner, P. O. Brown, A. Rein, *Proc. Natl. Acad. Sci. U.S.A.* 95, 7299 (1998).
21. J. T. Finch *et al.*, unpublished observations.
22. V. M. Vogt, in (2), pp. 27–70.
23. M. A. McClure, M. S. Johnson, D.-F. Feng, R. F. Doolittle, *Proc. Natl. Acad. Sci. U.S.A.* 83, 2469-2473 (1988).
24. Single-letter abbreviations for the amino acid residues are as follows: A, Ala; C, Cys; D, Asp; E, Glu; F, Phe; G, Gly; H, His; I, Ile; K, Lys; L, Leu; M, Met; N, Asn; P, Pro; Q, Gln; R, Arg; S, Ser; T, Thr; V, Val; W, Trp; and Y, Tyr.
25. We thank C. Hill for very helpful discussions on the relationship between viral cores and fullerene cones, D. Hobbs for refining the ChemDraw3D images of cones, G. Stubbs for a gift of tobacco mosiac virus, J. McCutcheon for the plasmid used to prepare ribosomal RNA, and K. Albertine and N. Chandler of the University of Utah Shared Electron Microscopy facility for their support and encouragement. Supported by grants from NIH and from the Huntsman Cancer Institute (to W.I.S.).

29 September 1998; accepted 17 November 1998

The Transcriptional Program in the Response of Human Fibroblasts to Serum

Vishwanath R. Iyer, Michael B. Eisen, Douglas T. Ross, Greg Schuler, Troy Moore, Jeffrey C. F. Lee, Jeffrey M. Trent, Louis M. Staudt, James Hudson Jr., Mark S. Boguski, Deval Lashkari, Dari Shalon, David Botstein, Patrick O. Brown*

The temporal program of gene expression during a model physiological response of human cells, the response of fibroblasts to serum, was explored with a complementary DNA microarray representing about 8600 different human genes. Genes could be clustered into groups on the basis of their temporal patterns of expression in this program. Many features of the transcriptional program appeared to be related to the physiology of wound repair, suggesting that fibroblasts play a larger and richer role in this complex multicellular response than had previously been appreciated.

The response of mammalian fibroblasts to serum has been used as a model for studying growth control and cell cycle progression (*1*). Normal human fibroblasts require growth factors for proliferation in culture; these growth factors are usually provided by fetal bovine serum (FBS). In the absence of growth factors, fibroblasts enter a nondividing state, termed G_0, characterized by low metabolic activity. Addition of FBS or purified growth factors induces proliferation of the fibroblasts; the changes in gene expression that accompany this proliferative response have been the subject of many studies, and the responses of dozens of genes to serum have been characterized.

We took a fresh look at the response of human fibroblasts to serum, using cDNA microarrays representing about 8600 distinct human genes to observe the temporal program of transcription that underlies this response. Primary cultured fibroblasts from human neonatal foreskin were induced to enter a quiescent state by serum deprivation for 48 hours and then stimulated by addition of medium containing 10% FBS (*2*). DNA microarray hybridization was used to measure the temporal changes in mRNA levels of 8613 human genes (*3*) at 12 times, ranging from 15 min to 24 hours after serum stimulation. The cDNA made from purified mRNA from each sample was labeled with the fluorescent dye Cy5 and mixed with a common reference probe consisting of cDNA made from purified mRNA from the quiescent

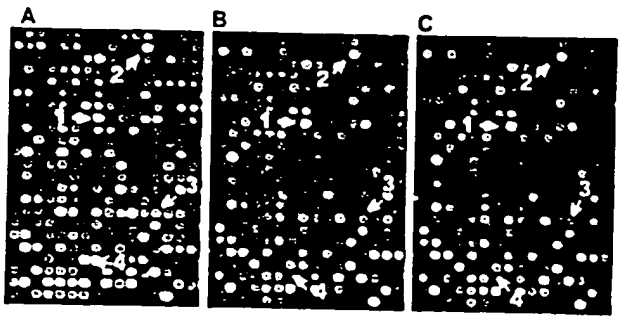

Fig. 1. The same section of the microarray is shown for three independent hybridizations comparing RNA isolated at the 8-hour time point after serum treatment to RNA from serum-deprived cells. Each microarray contained 9996 elements, including 9804 human cDNAs, representing 8613 different genes. mRNA from serum-deprived cells was used to prepare cDNA labeled with Cy3–deoxyuridine triphosphate (dUTP), and mRNA harvested from cells at different times after serum stimulation was used to prepare cDNA labeled with Cy5-dUTP. The two cDNA probes were mixed and simultaneously hybridized to the microarray. The image of the subsequent scan shows genes whose mRNAs are more abundant in the serum-deprived fibroblasts (that is, suppressed by serum treatment) as green spots and genes whose mRNAs are more abundant in the serum-treated fibroblasts as red spots. Yellow spots represent genes whose expression does not vary substantially between the two samples. The arrows indicate the spots representing the following genes: 1, protein disulfide isomerase-related protein P5; 2, IL-8 precursor; 3, EST AA057170; and 4, vascular endothelial growth factor.

V. R. Iyer and D. T. Ross, Department of Biochemistry, Stanford University School of Medicine, Stanford CA 94305, USA. M. B. Eisen and D. Botstein, Department of Genetics, Stanford University School of Medicine, Stanford CA 94305, USA. G. Schuler and M. S. Boguski, National Center for Biotechnology Information, Bethesda MD 20894, USA. T. Moore and J. Hudson Jr., Research Genetics, Huntsville, AL 35801, USA. J. C. F. Lee, D. Lashkari, D. Shalon, Incyte Pharmaceuticals, Fremont, CA 94555, USA. J. M. Trent, Laboratory of Cancer Genetics, National Human Genome Research Institute, National Institutes of Health, Bethesda, MD 20892, USA. L. M. Staudt, Metabolism Branch, Division of Clinical Sciences, National Cancer Institute, Bethesda, MD 20892, USA. P. O. Brown, Department of Biochemistry and Howard Hughes Medical Institute, Stanford University School of Medicine, Stanford CA 94305, USA.

*To whom correspondence should be addressed. E-mail: pbrown@cmgm.stanford.edu

culture (time zero) labeled with a second fluorescent dye, Cy3 (*4*). The color images of the hybridization results (Fig. 1) were made by representing the Cy3 fluorescent image as green and the Cy5 fluorescent image as red and merging the two color images.

Diverse temporal profiles of gene expression could be seen among the 8613 genes surveyed in this experiment (Fig. 2); many of these genes (about half) were unnamed expressed sequence tags (ESTs) (*5*). Although diverse patterns of expression were observed, the orderly choreography of the expression program became apparent when the results were analyzed by a clustering and display method developed in our laboratory for analyzing genome-wide gene expression data (*6*). An example of such an analysis, here applied to a subset of 517 genes whose expression changed substantially in response t serum (*7*), is shown in Fig. 2. The entire detailed data set underlying Fig. 2 is available as a tab-delimited table (in cluster order) at the *Science* Web site (www.sciencemag.org/feature/data/984559.shl). In addition, the entire, larger data set for the complete set of genes analyzed in this experiment can be found at a Web site maintained by our laboratory (genome-www.stanford.edu/serum) (*8*).

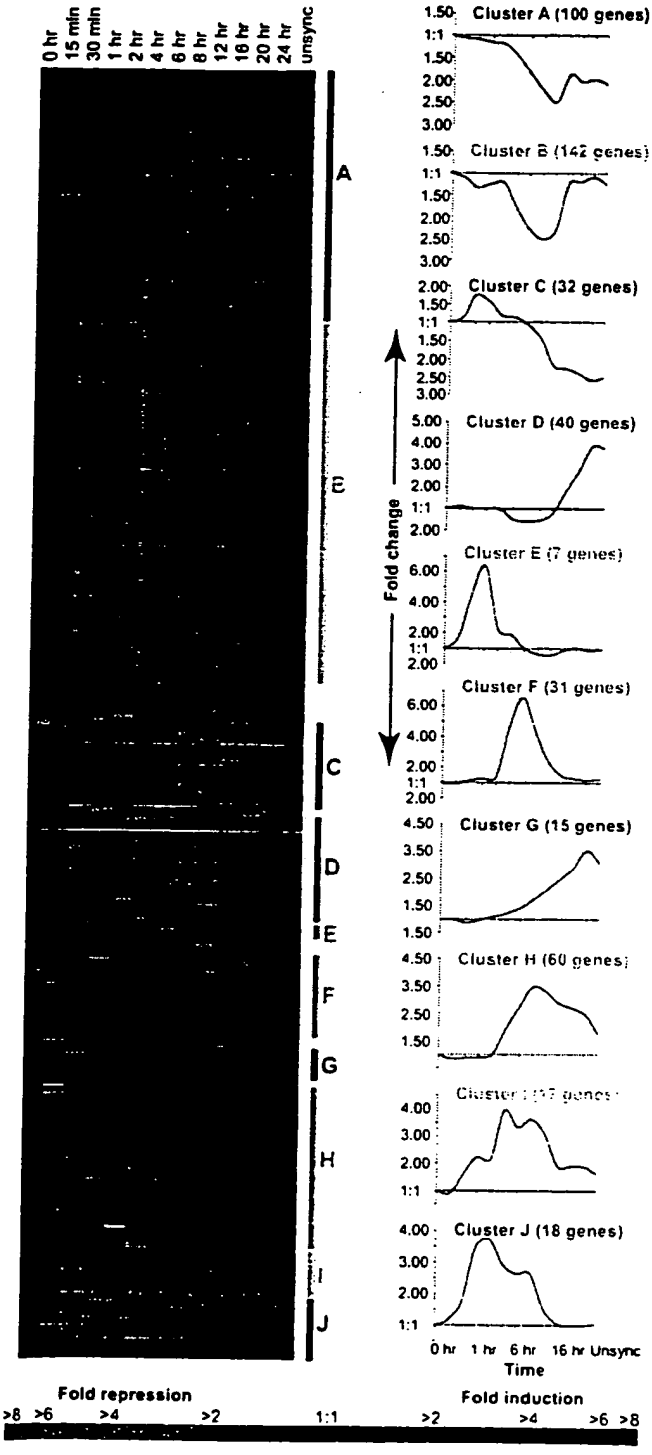

Fig. 2. Cluster image showing the different classes of gene expression profiles. Five hundred seventeen genes whose mRNA levels changed in response to serum stimulation were selected (*7*). This subset of genes was clustered hierarchically into groups on the basis of the similarity of their expression profiles by the procedure of Eisen *et al.* (*6*). The expression pattern of each gene in this set is displayed here as a horizontal strip. For each gene, the ratio of mRNA levels in fibroblasts at the indicated time after serum stimulation ("unsync" denotes exponentially growing cells) to its level in the serum-deprived (time zero) fibroblasts is represented by a color, according to the color scale at the bottom. The graphs show the average expression profiles for the genes in the corresponding "cluster" (indicated by the letters A to J and color coding). In every case examined, when a gene was represented by more than one array element, the multiple representations in this set were seen to have identical or very similar expression profiles, and the profiles corresponding to these independent measurements clustered either adjacent or very close to each other, pointing to the robustness of the clustering algorithm in grouping genes with very similar patterns of expression.

One measure of the reliability of the changes we observed is inherent in the expression profiles of the genes. For most genes whose expression levels changed, we could see a gradual change over a few time points, which thus effectively provided independent measurements for almost all of the observations. An additional check was provided by the inclusion of duplicate and, in a few cases, multiple array elements representing the same gene for about 5% of the genes included in this microarray. In addition, three independent hybridizations to different microarrays with mRNA samples from cells harvested 8 hours after serum addition showed good correlation (Fig. 1). As an independent test, we measured the expression levels of several genes using the TaqMan 5′ nuclease fluorigenic quantitative polymerase chain reaction (PCR) assay (*9*). The expression profiles of the genes, as measured by these two independent methods, were very similar (Fig. 3) (*10*).

The transcriptional response of fibroblasts to serum was extremely rapid. The immediate response to serum stimulation was dominated by genes that encode transcription factors and other proteins involved in signal transduction. The mRNAs for several genes [including c-FOS, JUN B, and mitogen-activated protein (MAP) kinase phosphatase–1 (MKP1)] were detectably induced within 15 min after serum stimulation (Fig. 4, A and B). Fifteen of the genes that were observed to be induced by serum encode known or suspected regulators of transcription (Fig. 4B). All but one were immediate-early genes—their induction was not inhibited by cycloheximide (*11*). This class of genes could be distinguished into those whose induction was transient (Fig. 2, cluster E) and those whose mRNA levels remained induced for much longer (Fig. 2, clusters I and J). Some features of the immediate response appeared to be directed at adaptation to the initiating signals. We observed a marked induction of mRNA encoding MKP1, a dual-specificity phosphatase that modulates the activity of the ERK1 and ERK2 MAP kinases (*12*). The coincidence of the peak of expression of genes in cluster E (Fig. 2) with that of MKP1 (Fig. 4A) suggests the possibility

that continued activity of the MAP kinase pathway is required to maintain induction of these genes but not of those with sustained expression (clusters I and J). The gene encoding a second member of the dual-specificity MAP kinase phosphatase family, known as dual-specificity protein phosphatase 6/pyst2, was induced later, at about 4 hours after serum stimulation. Genes encoding diverse other proteins with roles in signal transduction, ranging from cell-surface receptors [for example, the sphingosine 1–phosphate receptor (EDG-1), the vascular endothelial growth factor receptor, and the type II BMP receptor] to regulators of G-protein signaling (for example, NET1/p115 rho GEF) to DNA-binding transcription factors, were induced by serum (Fig. 4A).

The reprogramming of the regulatory circuits in response to serum involved not only induction of transcription factors but also reduced expression of many transcriptional regulators—some of which may play roles in maintaining the cells in G_0 or in priming them to react to wounding (Fig. 4C). Perhaps as a consequence of the historical focus on genes induced by serum stimulation of fibroblasts, the set of transcription factors whose expression diminished upon serum stimulation has been less well characterized.

Genes known or likely to be involved in controlling and mediating the proliferative response showed distinctive patterns of regulation. Several genes whose products inhibit progression of the cell-division cycle, such as p27 Kip1, p57 Kip2, and p18, were expressed in the quiescent fibroblasts and down-regulated before the onset of cell division. The nadir in the mRNA levels for these genes occurred between 6 and 12 hours after serum stimulation (Fig. 5A), coincident with the passage of the fibroblasts through G_1. The levels of the transcript encoding the WEE1-like protein kinase, which is believed to inhibit mitosis by phosphorylation of Cdc2, diminished between 4 and 8 to 12 hours after serum addition (Fig. 5A), well before the onset of M phase at around 16 hours, raising the possibility of an additional role for Wee1 in an earlier stage of the cell cycle or in regulating the G_0 to G_1 transition. Several genes induced in the first few hours after serum stimulation, such as the helix-loop-helix proteins ID2 and ID3 and EST AA016305, a gene with homology to G_1-S cyclins, are candidates for roles in promoting the exit from G_0.

Genes involved in mediating progression through the cell cycle were characterized by a distinctive pattern of expression (Fig. 2, cluster D), reflecting the coincidence of their expression with the reentry of the stimulated fibroblasts into the cell-division cycle. The stimulated fibroblasts replicated their DNA about 16 hours after serum treatment. This timing was reflected by the induction of mRNA encoding both subunits of ribonucleotide reductase and PCNA, the processivity factor for DNA polymerase epsilon and delta. Cyclin A, Cyclin B1, Cdc2, and CDC28 kinase, regulators of passage through the S phase and the transition from G_2 to M phase, were induced at about 16 to 20 hours after serum addition. The kinase in the Cyclin B1–CDK pair needs to be activated by phosphorylation. The gene encoding Cyclin-dependent kinase 7 (CDK7; a homolog of *Xenopus* MO15 cdk-activating kinase) was induced in parallel with the Cdc2 and Cdc28 kinases (Fig. 5A), suggesting a potential role for CDK7 in mediating M phase. DNA topoisomerase II α, required for chromosome segregation at mitosis; Mad2, a component of the spindle checkpoint that prevents completion of mitosis (anaphase) if chromosomes are not attached to the spindle; and the kinetochore protein CENP-F all showed a similar expression profile.

In the hours after the serum stimulus, one of the most striking features of the unfolding transcriptional program was the appearance of numerous genes with known roles in processes relevant to the physiology of wound healing. These included both genes involved in the direct role played by fibroblasts in remodeling of the clot and the extracellular matrix and, more notably, genes encoding proteins involved in intercellular signaling (Fig. 5). Genes induced in this program encode products that can (i) participate in the dynamic process of clotting, clot dissolution, and remodeling and perhaps contribute to hemostasis by promoting local vasoconstriction (for example, endothelin-1); (ii) promote chemotaxis and activation of neutrophils (for example, COX2) and recruitment and extravasation of monocytes and macrophages (for example, MCP1); (iii) promote chemotaxis and activation of T lymphocytes [for example, interleukin-8 (IL-8)] and B lymphocytes (for example, ICAM-1), thus providing both innate and antigen-specific defenses against wound infection and recruiting the phagocytic cells that will be required to clear out the debris during remodeling of the wound; (iv) promote angiogenesis and neovascularization (for example, VEGF) through newly forming tissue; (v) promote migration and proliferation of fibroblasts (for example, CTGF) and their differentiation into myofibroblasts (for example, Vimentin); and (vi) promote migration and proliferation of keratinocytes, leading to reepithelialization of the wound (for example, FGF7), and promote proliferation of melanocytes, perhaps contributing to wound hyperpigmentation (for example, FGF2).

Coordinated regulation of groups of genes whose products act at different steps in a common process was a recurring theme. For example, Furin, a prohormone-processing protease required for one of the processing steps in the generation of active endothelin, was induced in parallel with induction of the gene encoding the precursor of endothelin-1 (Fig. 5E) (*13*). Conversely, expression of CALLA/CD10, a membrane metalloprotease that degrades endothelin-1 and other peptide mediators of acute inflammation, was re-

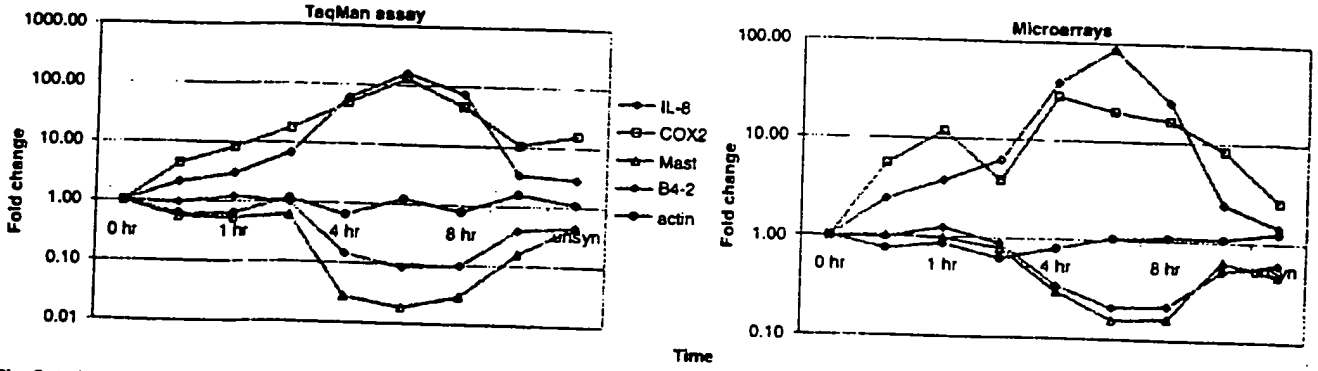

Fig. 3. Independent verification of microarray quantitation. Relative mRNA levels of the indicated genes (Mast, mast/stem cell growth factor receptor) were measured with the TaqMan 5′ nuclease fluorigenic quantitative PCR assay (*9*) (left) in the same samples that were used to prepare probes for microarray hybridizations (right). Data from the TaqMan analysis were normalized to mRNA concentrations and plotted relative to the level at time zero, so that the results could be compared with those from the microarray hybridizations. In general, quantitation with the two methods gave very similar results (*10*).

duced. A second example is provided by a set of five genes involved in the biosynthesis of cholesterol (Fig. 5I). The mRNAs encoding each of these enzymes showed sharply diminished expression beginning 4 to 6 hours after serum stimulation of fibroblasts. A likely explanation for the coordinated down-regulation of the cholesterol biosynthetic pathway is that serum provides cholesterol to fibroblasts through low-density lipoproteins, whereas in the absence of the cholesterol provided by serum, endogenous cholesterol biosynthesis in fibroblasts is required.

Many of the previously studied genes that we observed to be regulated in this program have no recognized role in any aspect of wound healing or fibroblast proliferation. Their identification in this study may therefore point to previously unknown aspects of these processes. A few selected genes in this group are shown in Fig. 5H. The stanniocalcin gene, for example (Fig. 5H), encodes a secreted protein without a clearly identified function in human cells (*14*, *15*). Its induction in serum-stimulated fibroblasts suggests the possibility that it may play a role in the wound-healing process, perhaps serving as a signal in mediating inflammation or angiogenesis.

One of the most important results of this exploration was the discovery of over 200 previously unknown genes whose expression was regulated in specific temporal patterns during the response of fibroblasts to serum. For example, 13 of the 40 genes in cluster D (Fig. 2) have descriptive names that reflect their putative function. Nine of these 13 genes (69%) encode proteins that play roles in cell cycle progression, particularly in DNA replication and the G_2-M transition. This enrichment for cell cycle–related genes suggests that some of the unnamed genes in this cluster—for example, EST W79311 and EST R13146, neither of which have sequence similarity to previously characterized genes—may represent previously unknown genes involved in this part of the cell cycle. Similarly, a remarkable fraction of genes that were grouped into cluster F on the basis of their expression profiles encoded proteins involved in intercellular signaling (Fig. 2), suggesting that a similar role should be considered for the many unnamed genes in this cluster. A disproportionately large fraction of the genes whose transcription diminished upon serum stimulation were unnamed ESTs.

Our intention was to use this experiment as a model to study the control of the transition

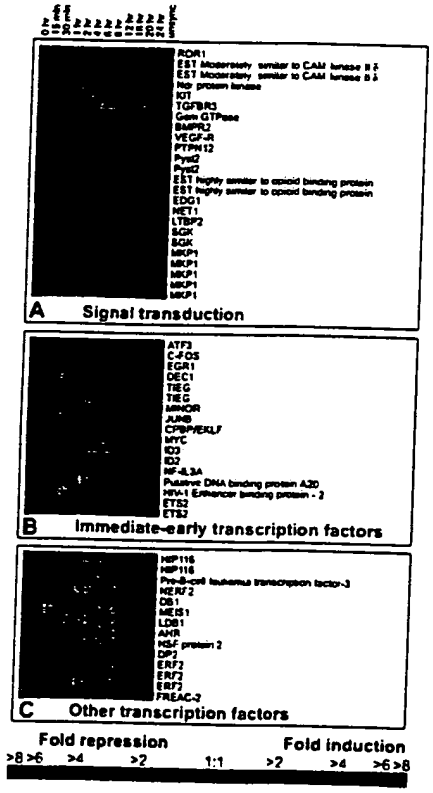

Fig. 4. "Reprogramming" of fibroblasts. Expression profiles of genes whose function is likely to play a role in the reprogramming phase of the response are shown with the same representation as in Fig. 2. In the cases in which a gene was represented by more than one element in the microarray, all measurements are shown. The genes were grouped into categories on the basis of our knowledge of their most likely role. Some genes with pleiotropic roles were included in more than one category.

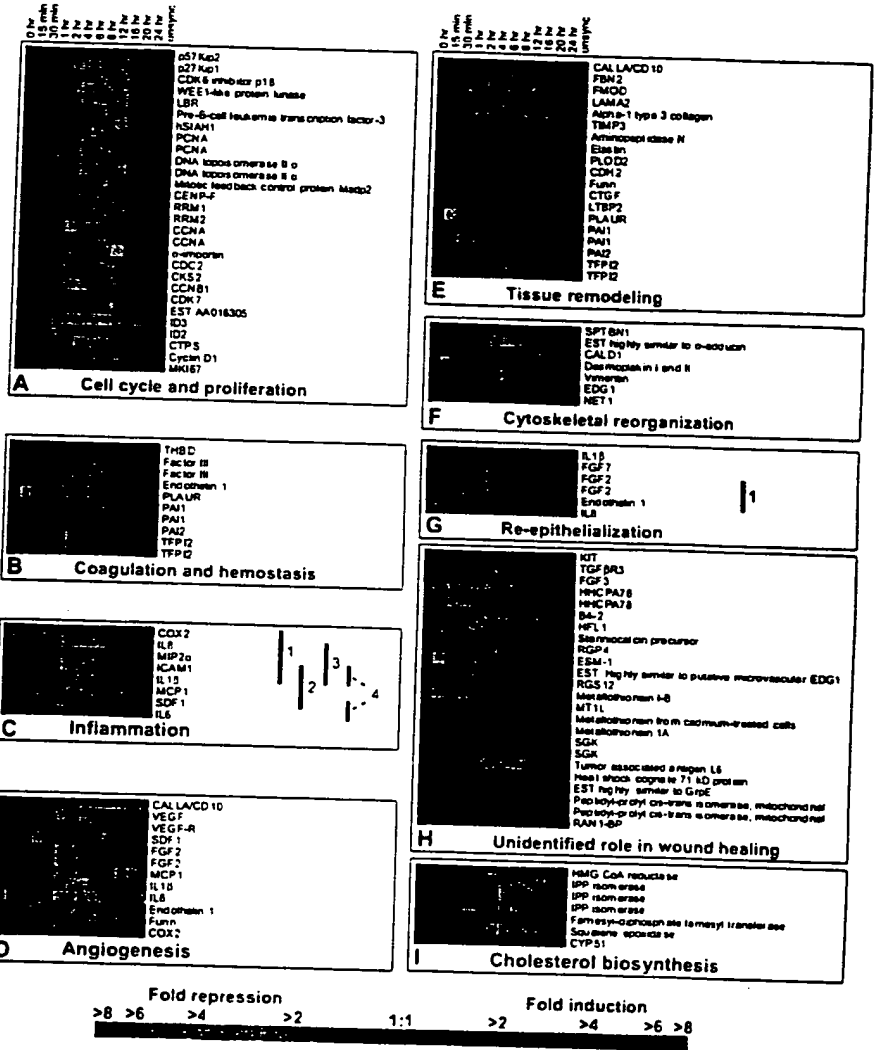

Fig. 5. The transcriptional response to serum suggests a multifaceted role for fibroblasts in the physiology of wound healing. The features of the transcriptional program of fibroblasts in response to serum stimulation that appear to be related to various aspects of the wound-healing process and fibroblast proliferation are shown with the same convention for representing changes in transcript levels as was used in Figs. 2 and 4. (A) Cell cycle and proliferation, (B) coagulation and hemostasis, (C) inflammation, (D) angiogenesis, (E) tissue remodeling, (F) cytoskeletal reorganization, (G) reepithelialization, (H) unidentified role in wound healing, and (I) cholesterol biosynthesis. The numbers in (C) and (G) refer to genes whose products serve as signals to neutrophils (C1), monocytes and macrophages (C2), T lymphocytes (C3), B lymphocytes (C4), and melanocytes (G1).

from G_0 to a proliferating state. However, one of the defining characteristics of genome-scale expression profiling experiments is that the examination f so many diverse genes opens a window on all the processes that actually occur and not merely the single process one intended to observe. Serum, the soluble fraction of clotted blood, is normally encountered by cells in vivo in the context of a wound. Indeed, the expression program that we observed in response to serum suggests that fibroblasts are programmed to interpret the abrupt exposure to serum not as a general mitogenic stimulus but as a specific physiological signal, signifying a wound. The proliferative response that we originally intended to study appeared to be part of a larger physiological response of fibroblasts to a wound. Other features of the transcriptional response to serum suggest that the fibroblast is an active participant in a conversation among the diverse cells that work together in wound repair, interpreting, amplifying, modifying, and broadcasting signals controlling inflammation, angiogenesis, and epithelial regrowth during the response to an injury.

We recognize that these in vitro results almost certainly represent a distorted and incomplete rendering of the normal physiological response of a fibroblast to a wound. Moreover, only the responses elicited directly by exposure of fibroblasts to serum were examined. The subsequent signals from other cellular participants in the normal wound-healing process would certainly provoke further evolution of the transcriptional program in fibroblasts at the site of a wound, which this experiment cannot reveal. Nevertheless, we believe that the picture that emerged strongly suggests a much larger and richer role for the fibroblast in the orchestration of this important physiological process than had previously been suspected.

References and Notes

1. J. A. Winkles, *Prog. Nucleic Acid Res. Mol. Biol.* **58**, 41 (1998).
2. A normal human diploid fibroblast cell line derived from foreskin (ATCC CRL 2091) in passage 8 was used in these experiments. The protocol followed for growth arrest and stimulation was essentially that of (*16*) and (*17*). Cells were grown to about 60% confluence in 15-cm petri dishes in Dulbecco's minimum essential medium containing glucose (1 g/liter), the antibiotics penicillin and streptomycin, and 10% (by vol) FBS (Hyclone) that had been previously heat inactivated at 56°C for 30 min. The cells were then washed three times with the same medium lacking FBS, and low-serum medium (0.1% FBS) was added to the plates. After a 48-hour incubation, the medium was replaced with fresh medium containing 10% FBS. mRNA was isolated from several plates of cells harvested before serum stimulation; this mRNA served as the serum-starved or time-zero reference sample. Cells were harvested from batches of plates at 11 subsequent intervals (15 min, 30 min, 1, 2, 4, 6, 8, 12, 16, 20, and 24 hours) after the addition of serum. mRNA was also isolated from exponentially growing fibroblasts (not subjected to serum starvation). mRNA was isolated with the FastTrack mRNA isolation kit (Invitrogen), which involves lysis of the cells on the plate. The growth medium was removed, and the cells were quickly washed with phosphate-buffered saline at room temperature. The lysis buffer was added to the plate, transferred to tubes, and frozen in liquid nitrogen. Subsequent steps were performed according to the kit manufacturer's protocols.
3. The National Center for Biotechnology Information maintains the UniGene database as a resource for partitioning human sequences contained in GenBank into clusters representing distinct transcripts or genes (*18, 19*). At the time this work began, this database contained about 40,000 such clusters. We selected a subset of 10,000 of these UniGene clusters for inclusion on gene expression microarrays. UniGene clusters were included only if they contained at least one clone from the I.M.A.G.E. human cDNA collection (*20*), so that a physical clone could easily be obtained (all I.M.A.G.E. clones are available commercially from a number of vendors). We attempted to include as complete as possible a set of the "named" human genes (about 4000) and genes that appeared to be closely related to named genes in other organisms (about an additional 2000). The remaining 4000 clones were chosen from among the "anonymous" UniGene clusters on the basis of inclusion on the human transcript map (www.ncbi.nlm.nih.gov/SCIENCE96/) and the lack of apparent homology to any other genes in the selected set. A physical clone representing each of the selected genes was obtained from Research Genetics. This "10K set" is included in a more recent "15K set" described at www.nhgri.nih.gov/DIR/LCG/15K/HTML/p15Ktop.html. Of these clones, 472 are absent from the current edition of UniGene and were presumed to be distinct genes. The remainders represent 8141 distinct clusters, or human genes, in UniGene. These clones, thus presumed to represent 8613 different genes, were used to print microarrays according to methods described previously (*21, 22*).
4. One microgram of mRNA was used for making fluorescently labeled cDNA probes for hybridizing to the microarrays, with the protocol described previously (*23*). mRNA from the large batch of serum-starved cells was used to make cDNA labeled with Cy3. The Cy3-labeled cDNA from this batch of serum-starved cells served as the common reference probe in all hybridizations. mRNA samples from cells harvested immediately before serum stimulation, at intervals after serum stimulation, and from exponentially growing cells were used to make cDNA labeled with Cy5. Ten micrograms of yeast tRNA, 10 μg of polydeoxyadenylic acid, and 20 μg of human CoT1 DNA (Gibco-BRL) were added to the mixture of labeled probes in a solution containing 3× standard saline citrate (SSC) and 0.3% SDS and allowed to prehybridize at room temperature for 30 min before the probe was added to the surface of the microarray. Hybridizations, washes, and fluorescent scans were performed as described previously (*23, 24*). All measurements, totaling more than 180,000 differential expression measurements, were stored in a computer database for analysis and interpretation.
5. The nominal identities of a number of cDNAs (currently about 3750) on the microarray were verified by sequencing. The clones that were sequenced included many of the genes whose expression changed substantially upon serum stimulation, as well as a large number of genes whose expression did not change substantially in the course of this experiment. About 85% of the clones on the current version of this microarray that were checked by resequencing were correctly identified. In all the figures, gene names or EST numbers are given only for those genes on the microarrays whose identities were reconfirmed by resequencing. In the cases where a human gene has more than one name in the literature, we have tried to use the name that is most evocative of its presumed role in this context. The remainder of the clones have been assigned a temporary identification number (format: SID######) and a putative identity pending sequence verification. The correct identities of these genes will be posted at our Web site (genome-www.stanford.edu/serum) as they are confirmed by resequencing.
6. M. B. Eisen, P. T. Spellman, P. O. Brown, D. Botstein, *Proc. Natl. Acad. Sci. U.S.A.* **95**, 14863 (1998).
7. Genes were selected for this analysis if either (i) their expression level deviated from that in quiescent fibroblasts by at least a factor of 2.20 in at least two of the samples from serum-stimulated cells or (ii) the standard deviation for the set of 13 values of $\log_2$(expression ratio) measured for the gene in this time course exceeded 0.7. In addition, observations in which the pixel-by-pixel correlation coefficients for the Cy3 and Cy5 fluorescence signals measured in a given array element were less than 0.6 were excluded. This selection criteria yielded a computationally manageable number of genes while minimizing the number of genes that were included because of noise in the data.
8. A more complete analysis and interpretation of the results of this experiment, as well as a searchable database, can be found at genome-www.stanford.edu/serum
9. K. J. Livak, S. J. Flood, J. Marmaro, W. Giusti, K. Deetz, *PCR Methods. Appl.* **4**, 357 (1995).
10. The apparent dip in the profile of COX2 at the 2-hour time point in the microarray data appears to result from a localized area of low intensity on the corresponding array scan resulting in an underestimation of the expression ratio. The expression ratios measured for mast/stem cell growth factor receptor are somewhat lower in the microarray data. This discrepancy is probably a consequence of the conservative background subtraction method used for quantitating the signal intensities on the array scans (*23*). The sequences of the PCR primer pairs (5′ to 3′) that were used are as follows: COX2, CCGTGGCTCTCTTGGCAG and CTAAGTTCTTTAGCACTCCTTGGCA; IL-8, CGATGCTGTGGAGCTGTATC and CCATGGTTTCACCAAAGATG; mast/stem cell factor receptor, ACAGAAGCCCGTGGTAGACC and GAGGCTGGGAGGAGGAAG; B4-2, AAACCCCCCTCAGGAAAGAC and CCATGAACAAGCTGGCCAT; and actin, AGTACTCCGTGTGGATCGGC and GCTGATCCACATCTGCTGGA
11. V. R. Iyer *et al.*, unpublished data. The gene expression data for the early time points in the presence of cycloheximide will be available at our Web site (genome-www.stanford.edu/serum)
12. T. Hunter, *Cell* **80**, 225 (1995).
13. J. Leppaluoto and H. Ruskoaho, *Ann. Med.* **24**, 153 (1992).
14. A. C. Chang *et al.*, *Mol. Cell. Endocrinol.* **112**, 241 (1995).
15. K. L. Madsen *et al.*, *Am. J. Physiol.* **274**, G96 (1998).
16. W. Krek and J. A. DeCaprio, *Methods Enzymol.* **254**, 114 (1995).
17. R. A. Tobey, J. G. Valdez, H. A. Crissman, *Exp. Cell Res.* **179**, 400 (1988).
18. M. S. Boguski and G. D. Schuler, *Nature Genet.* **10**, 369 (1995).
19. G. D. Schuler, *J. Mol. Med.* **75**, 694 (1997).
20. G. Lennon, C. Auffray, M. Polymeropoulos, M. B. Soares, *Genomics* **33**, 151 (1996).
21. I.M.A.G.E. clones were amplified by PCR in 96-well format with amino-linked primers at the 5′ end. Purified PCR products were suspended at a concentration of ~0.5 mg/ml in 3× SSC, and ~5 ng of each product was arrayed onto coated glass by means of procedures similar to those described previously (*22*). A total of 9996 elements were arrayed onto an area of 1.8 cm by 1.8 cm with the elements spaced 175 μm apart. The microarrays were then postprocessed to fix the DNA to the glass surface before hybridization with a procedure similar to previously described methods (*22*).
22. M. Schena, D. Shalon, R. W. Davis, P. O. Brown, *Science* **270**, 467 (1995).
23. J. L. DeRisi, V. R. Iyer, P. O. Brown, *ibid.* **278**, 680 (1997).
24. J. DeRisi *et al.*, *Nature Genet.* **14**, 457 (1996).

25. We thank E. Chung for help with sequencing, A. Alizadeh for help with sequence verification, K. Ranade for advice on the TaqMan assay, and J. DeRisi and other members of the P.O.B. and D.B. labs for discussions. Supported by a grant from the National Human Genome Research Institute (NHGRI) (HG00450) and the National Cancer Institute (NIH CA 77097). V.R.I. was supported in part by an Institutional Training Grant in Genome Sciences (T32 HG00044) from the NHGRI. M.B.E. is an Alfred E. Sloan Foundation Postdoctoral Fellow in Computational Molecular Biology, and D.T.R is a Walter and Idun Berry Fellow. P.O.B. is an Associate Investigator of the Howard Hughes Medical Institute.

13 August 1998; accepted 13 November 1998

Systematic variation in gene expression patterns in human cancer cell lines

Douglas T. Ross[1], Uwe Scherf[5], Michael B. Eisen[2], Charles M. Perou[2], Christian Rees[2], Paul Spellman[2], Vishwanath Iyer[1], Stefanie S. Jeffrey[3], Matt Van de Rijn[4], Mark Waltham[5], Alexander Pergamenschikov[2], Jeffrey C.F. Lee[6], Deval Lashkari[7], Dari Shalon[6], Timothy G. Myers[8], John N. Weinstein[5], David Botstein[2] & Patrick O. Brown[1,9]

We used cDNA microarrays to explore the variation in expression of approximately 8,000 unique genes among the 60 cell lines used in the National Cancer Institute's screen for anti-cancer drugs. Classification of the cell lines based solely on the observed patterns of gene expression revealed a correspondence to the ostensible origins of the tumours from which the cell lines were derived. The consistent relationship between the gene expression patterns and the tissue of origin allowed us to recognize outliers whose previous classification appeared incorrect. Specific features of the gene expression patterns appeared to be related to physiological properties of the cell lines, such as their doubling time in culture, drug metabolism or the interferon response. Comparison of gene expression patterns in the cell lines to those observed in normal breast tissue or in breast tumour specimens revealed features of the expression patterns in the tumours that had recognizable counterparts in specific cell lines, reflecting the tumour, stromal and inflammatory components of the tumour tissue. These results provided a novel molecular characterization of this important group of human cell lines and their relationships to tumours *in vivo*.

Introduction

Cell lines derived from human tumours have been extensively used as experimental models of neoplastic disease. Although such cell lines differ from both normal and cancerous tissue, the inaccessibility of human tumours and normal tissue makes it likely that such cell lines will continue to be used as experimental models for the foreseeable future. The National Cancer Institute's Developmental Therapeutics Program (DTP) has carried out intensive studies of 60 cancer cell lines (the NCI60) derived from tumours from a variety of tissues and organs[1–4]. The DTP has assessed many molecular features of the cells related to cancer and chemotherapeutic sensitivity, and has measured the sensitivities of these 60 cell lines to more than 70,000 different chemical compounds, including all common chemotherapeutics (http://dtp.nci.nih.gov). A previous analysis of these data revealed a connection between the pattern of activity of a drug and its method of action. In particular, there was a tendency for groups of drugs with similar patterns of activity to have related methods of action[3,5–7].

We used DNA microarrays to survey the variation in abundance of approximately 8,000 distinct human transcripts in these 60 cell lines. Because of the logical connection between the function of a gene and its pattern of expression, the correlation of gene expression patterns with the variation in the phenotype of the cell can begin the process by which the function of a gene can be inferred. Similarly, the patterns of expression of known genes can reveal novel phenotypic aspects of the cells and tissues studied[8–10]. Here we present an analysis of the observed patterns of gene expression and their relationship to phenotypic properties of the 60 cell lines. The accompanying report[11] explores the relationship between the gene expression patterns and the drug sensitivity profiles measured by the DTP. The assessment of gene expression patterns in a multitude of cell and tissue types, such as the diverse set of cell lines we studied here, under diverse conditions *in vitro* and *in vivo*, should lead to increasingly detailed maps of the human gene expression program and provide clues as to the physiological roles of uncharacterized genes[11–16]. The databases, plus tools for analysis and visualization of the data, are available (http://genome-www.stanford.edu/nci60 and http://discover.nci.nih.gov).

Results

We studied gene expression in the 60 cell lines using DNA microarrays prepared by robotically spotting 9,703 human cDNAs on glass microscope slides[17,18]. The cDNAs included approximately 8,000 different genes: approximately 3,700 represented previously characterized human proteins, an additional 1,900 had homologues in other organisms and the remaining 2,400 were identified only by ESTs. Due to ambiguity of the identity of the cDNA clones used in these studies, we estimated that approximately 80% of the genes in these experiments were correctly identified. The identities of approximately 3,000 cDNAs

Departments of [1]Biochemistry, [2]Genetics, [3]Surgery and [4]Pathology, Stanford University School of Medicine, Stanford, California, USA. [5]Laboratory of Molecular Pharmacology, Division of Basic Sciences, National Cancer Institute, National Institutes of Health, Bethesda, Maryland, USA. [6]Incyte Pharmaceuticals, Fremont, California, USA. [7]Genometrix Inc., The Woodlands, Texas, USA. [8]Information Technology Branch, Developmental Therapeutics Program, Division of Cancer Treatment and Diagnosis, National Cancer Institute, National Institutes of Health, Rockville, Maryland, USA. [9]Howard Hughes Medical Institute, Stanford University School of Medicine, Stanford, California, USA. Correspondence should be addressed to P.O.B. (e-mail: pbrown@cmgm.stanford.edu) or J.N.W. (e-mail: Weinstein@dtpax2.ncifcrf.gov).

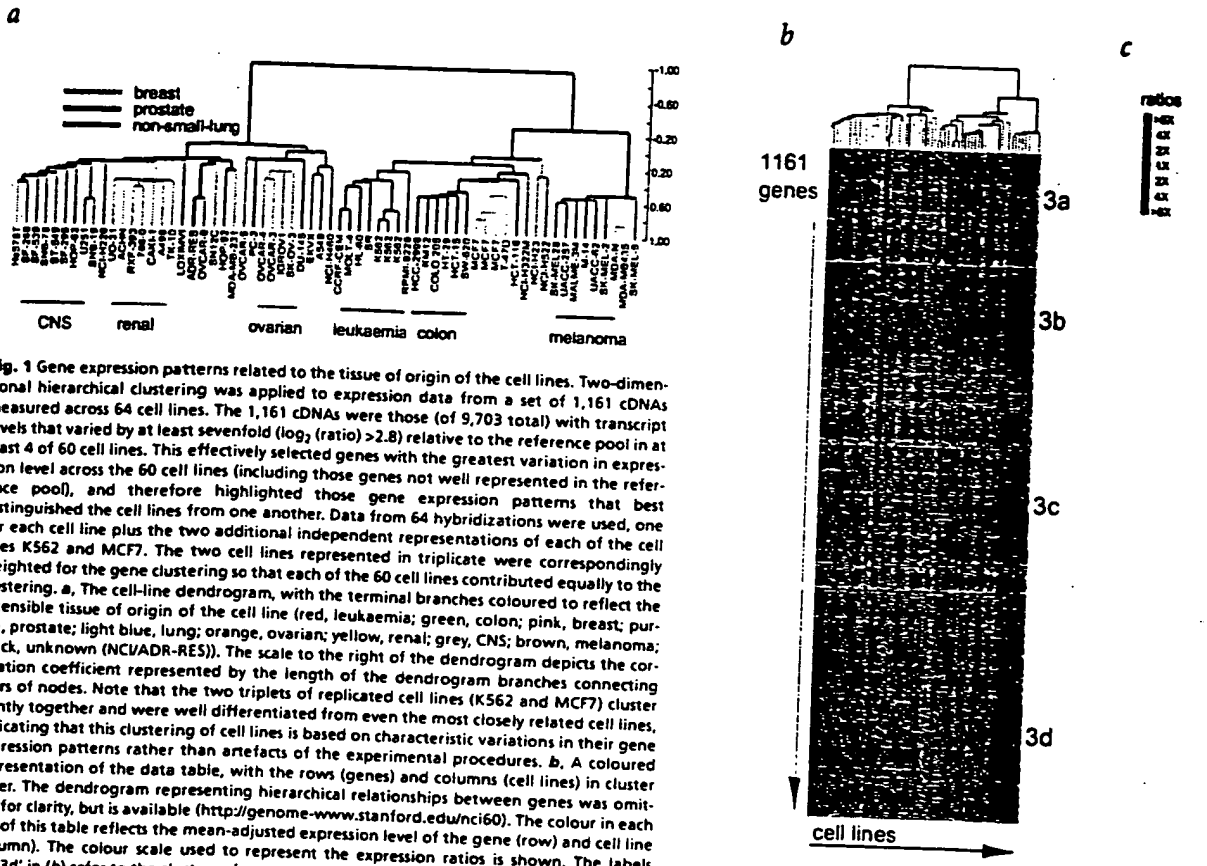

Fig. 1 Gene expression patterns related to the tissue of origin of the cell lines. Two-dimensional hierarchical clustering was applied to expression data from a set of 1,161 cDNAs measured across 64 cell lines. The 1,161 cDNAs were those (of 9,703 total) with transcript levels that varied by at least sevenfold ($\log_2$ (ratio) >2.8) relative to the reference pool in at least 4 of 60 cell lines. This effectively selected genes with the greatest variation in expression level across the 60 cell lines (including those genes not well represented in the reference pool), and therefore highlighted those gene expression patterns that best distinguished the cell lines from one another. Data from 64 hybridizations were used, one for each cell line plus the two additional independent representations of each of the cell lines K562 and MCF7. The two cell lines represented in triplicate were correspondingly weighted for the gene clustering so that each of the 60 cell lines contributed equally to the clustering. *a*, The cell-line dendrogram, with the terminal branches coloured to reflect the ostensible tissue of origin of the cell line (red, leukaemia; green, colon; pink, breast; purple, prostate; light blue, lung; orange, ovarian; yellow, renal; grey, CNS; brown, melanoma; black, unknown (NCI/ADR-RES)). The scale to the right of the dendrogram depicts the correlation coefficient represented by the length of the dendrogram branches connecting pairs of nodes. Note that the two triplets of replicated cell lines (K562 and MCF7) cluster tightly together and were well differentiated from even the most closely related cell lines, indicating that this clustering of cell lines is based on characteristic variations in their gene expression patterns rather than artefacts of the experimental procedures. *b*, A coloured representation of the data table, with the rows (genes) and columns (cell lines) in cluster order. The dendrogram representing hierarchical relationships between genes was omitted for clarity, but is available (http://genome-www.stanford.edu/nci60). The colour in each cell of this table reflects the mean-adjusted expression level of the gene (row) and cell line (column). The colour scale used to represent the expression ratios is shown. The labels '3a–3d' in (*b*) refer to the clusters of genes shown in detail in Fig. 3.

from these experiments have been sequence-verified, including all of those referred to here by name.

Each hybridization compared Cy5-labelled cDNA reverse transcribed from mRNA isolated from one of the cell lines with Cy3-labelled cDNA reverse transcribed from a reference mRNA sample. This reference sample, used in all hybridizations, was prepared by combining an equal mixture of mRNA from 12 of the cell lines (chosen to maximize diversity in gene expression as determined primarily from two-dimensional gel studies[2]). By comparing cDNA from each cell line with a common reference, variation in gene expression across the 60 cell lines could be inferred from the observed variation in the normalized Cy5/Cy3 ratios across the hybridizations.

To assess the contribution of artefactual sources of variation in the experimentally measured expression patterns, K562 and MCF7 cell lines were each grown in three independent cultures, and the entire process was carried out independently on mRNA extracted from each culture. The variance in the triplicate fluorescence ratio measurements approached a minimum when the fluorescence signal was greater than approximately 0.4% of the measurable total signal dynamic range above background in either channel of the hybridization. We selected the subset of spots for which significant signal was present in both the numerator and denominator of the ratios by this criterion to identify the best-measured spots. The pair-wise correlation coefficients for the triplicates of the set of genes that passed this quality control level (6,992 spots included for the MCF7 samples and 6,161 spots for K562) ranged from 0.83 to 0.92 (for graphs and details, see http://genome-www.stanford.edu/nci60).

To make the orderly features in the data more apparent, we used a hierarchical clustering algorithm[19,20] and a pseudo-colour visualization matrix[3,21]. The object of the clustering was to group cell lines with similar repertoires of expressed genes and to group genes whose expression level varied among the 60 cell lines in a similar manner. Clustering was performed twice using different subsets of genes to assess the robustness of the analysis. In one case (Fig. 1), we concentrated on those genes that showed the most variation in expression among the 60 cell lines (1,167 total). A second analysis (Fig. 2) included all spots that were thought to be well measured in the reference set (6,831 spots).

Gene expression patterns related to the histologic origins of the cell lines

The most notable property of the clustered data was that cell lines with common presumptive tissues of origin grouped together (Figs 1*a* and 2). Cell lines derived from leukaemia, melanoma, central nervous system, colon, renal and ovarian tissue were clustered into independent terminal branches specific to their respective organ types with few exceptions. Cell lines derived from non-small lung carcinoma and breast tumours were distributed in multiple different terminal branches suggesting that their gene expression patterns were more heterogeneous.

Many of these coherent cell line clusters were distinguished by the specific expression of characteristic groups of genes (Fig. 3*a–d*). For example, a cluster of approximately 90 genes was highly expressed in the melanoma-derived lines (Fig. 3*c*). This set was enriched for genes with known roles in melanocyte biology, including tyrosinase and dopachrome tautomerase (TYR and DCT; two subunits of an enzyme complex involved in melanin synthesis[22]), MART1 (MLANA; which is being investigated as a target for immunotherapy of melanoma[23]) and S100-β (S100B; which has been used as an antigenic marker in the diagnosis of

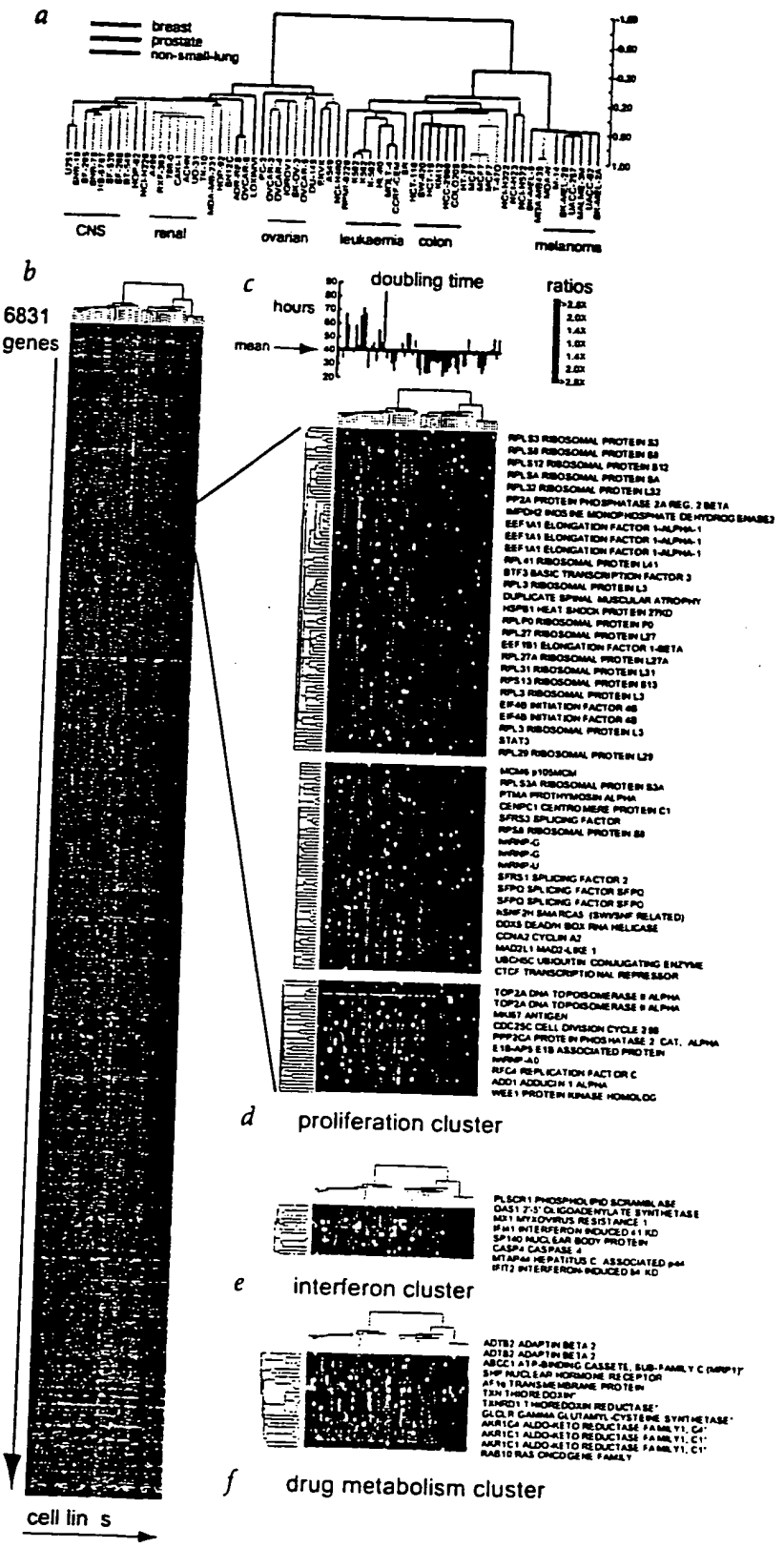

Fig. 2 Gene expression patterns related to other cell-line phenotypes. *a*, We applied two-dimensional hierarchical clustering to expression data from a set of 6,831 cDNAs measured across the 64 cell lines. The 6,831 cDNAs were those with a minimum fluorescence signal intensity of approximately 0.4% of the dynamic range above background in the reference channel in each of the six hybridizations used to establish reproducibility. This effectively selected those spots that provided the most reliable ratio measurements and therefore identified a subset of genes useful for exploring patterns comprised of those whose variation in expression across the 60 cell lines was of moderate magnitude. *b*, Cluster-ordered data table. *c*, Doubling time of cell lines. Cell lines are given in cluster order. Values are plotted relative to the mean. Doubling times greater than the mean are shown in green, those with doubling time less than the mean are shown in red. *d*, Three related gene clusters that were enriched for genes whose expression level variation was correlated with cell line proliferation rate. Each of the three gene clusters (clustered solely on the basis of their expression patterns) showed enrichment for sets of genes involved in distinct functional categories (for example, ribosomal genes versus genes involved in pre-RNA splicing). *e*, Gene cluster in which all characterized and sequence-verified cDNAs encode genes known to be regulated by interferons. *f*, Gene cluster enriched for genes that have been implicated in drug metabolism (indicated by asterisks). A further property of the gene clustering evident here and in Fig. 2 is the strong tendency for redundant representations of the same gene to cluster immediately adjacent to one another, even within larger groups of genes with very similar expression patterns. In addition to illustrating the reproducibility and consistency of the measurements, and providing independent confirmation of many of our measurements, this property also demonstrates that these, and probably all, genes have nearly unique patterns of variation across the 60 cell lines. If this were not the case, and multiple genes had identical patterns of variation, we would not expect to be able to distinguish, by clustering on the basis of expression variation, duplicate copies of individual genes from the other genes with identical expression patterns.

melan ma). LOXIMVI, the seventh line designated as melanoma in the NCI60, did not show this characteristic pattern. Although isolated fr m a patient with melanoma, LOXIMVI has previously been noted t lack melanin and other markers useful for identificati n of melanoma cells[1].

Paradoxically, two related cell lines (MDA-MB435 and MDA-N), which were derived from a single patient with breast cancer and have been conventionally regarded as breast cancer cell lines, shared expression of the genes associated with melanoma. MDA-MB435 was isolated from a pleural effusion in a patient with metastatic ductal adenocarcinoma of the breast[24,25]. It remains possible that the origin of the cell line was a breast cancer, and that its gene expression pattern is related to the neuroendocrine features of some breast cancers[26]. But our results suggest that this cell line may have originated from a melanoma, raising the possibility that the patient had a co-existing occult melanoma.

The higher-level organization of the cell-line tree—in which groups span cell lines from different tissue types—also reflected shared biological properties of the tissues from which the cell lines were derived. The carcinoma-derived cell lines were divided into major branches that separated those that expressed genes characteristic of epithelial cells from those that expressed genes more typical of stromal cells. A cluster of genes is shown (Fig. 3*b*) that is most strongly expressed in cell lines derived from colon carcinomas, six of seven ovarian-derived cell lines and the two breast cancer lines positive for the oestrogen receptor. The named genes in this cluster have been implicated in several aspects of epithelial cell biology[27]. The cluster was enriched for genes whose products are known to localize to the basolateral membrane of epithelial cells, including those encoding components of adherens complexes (for example, desmoplakin (DSP), periplakin (PPL) and plakoglobin (JUP)), an epithelial-expressed cell-cell adhesion molecule (M4S1) and a sodium/hydrogen ion exchanger[28–31] (SLC9A1). It also contained genes that encode putative transcriptional regulators of epithelial morphogenesis, a human homologue of a *Drosophila melanogaster* epithelial-expressed tumour suppressor (LLGL1) and a homeobox gene thought to control calcium-mediated adherence in epithelial cells[32,33] (MSX2).

In contrast, a separate, major branch of the cell-line dendrogram (Fig. 1*a*) included all glioblastoma-derived cell lines, all renal-cell-carcinoma–derived cell lines and the remaining carcinoma-derived lines. The characteristic set of genes expressed in this cluster included many whose products are involved in stromal cell functions (Fig. 3*d*). Indeed, the two cell lines originally described as 'sarcoma-like' in appearance (Hs578T, breast carcinosarcoma, and SF539, gliosarcoma) expressed most of these genes[34,35]. Although no single gene was uniformly characteristic of this cluster, each cell line showed a distinctive pattern of expression of genes encoding proteins with roles in synthesis or modification of the extracellular matrix (for example, caldesmon (CALD1), cathepsins, thrombospondin (THBS), lysyl oxidase (LOX) and collagen subtypes). Although the ovarian and most non-small-cell-lung–derived carcinomas expressed genes characteristic of both epithelial cells and stromal cells, they probably clustered with the CNS and renal cell carcinomas in this analysis because genes characteristically expressed in stromal cells were more abundantly represented in this gene set.

Physiol gical variation reflect d in gene expressi n patt rns

A cluster diagram of 6,831 genes (Fig. 2) is useful for exploring clusters of genes whose variation in mRNA levels was not obviously attributable to cell or tissue type. We identified some gene clusters that were enriched for genes involved in specific cellular processes; the variation in their expression levels may reflect corresponding differences in activity of these processes in the cell lines. For example, a cluster f 1,159 genes (Fig. 2*a*) included many whose products are necessary for progression through the cell cycle (such as CCNA1, MCM106 and MAD2L1), RNA processing and translation machinery (such as RNA helicases, hnRNPs and translation elongation factors) and traditional pathologic markers used to identify proliferating cells (MKI67). Within this large cluster were smaller clusters enriched for genes with more specialized roles. One cluster was highly enriched for numerous ribosomal genes, whereas another was more enriched for genes encoding RNA-splicing factors. The variation in expression of these ribosomal genes was significantly correlated with variation in the cell doubling time (correlation coefficient of 0.54), supporting the notion that the genes in this cluster were regulated in relation to cell proliferation rate or growth rate in these cell lines.

In a smaller gene cluster (Fig. 2*d*), all of the named genes were previously known to be regulated by interferons[13,36]. Additional groups of interferon-regulated genes showed distinct patterns of expression (data not shown), suggesting that the NCI60 cell lines exhibited variation in activity of interferon-response pathways, which was reflected in gene expression patterns[36].

Another cluster (Fig. 2*e*) contained several genes encoding proteins with possible interrelated roles in drug metabolism, including glutamate-cysteine ligase (GLCLC, the enzyme responsible for the rate limiting step of glutathione synthesis), thioredoxin (TXN) and thioredoxin reductase (TXNRD1; enzymes involved in regulating redox state in cells), and MRP1 (a drug transporter known to efficiently transport glutathione-conjugated compounds[37]). The elevated expression of this set of genes in a subset of these cell lines may reflect selection for resistance to chemotherapeutics.

Cell lines facilitate interpretation of gene expression patterns in complex clinical samples

Like many other types of cancer, tumours of the breast typically have a complex histological organization, with connective tissue and leukocytic infiltrates interwoven with tumour cells. To explore the possibility that variation in gene expression in the tumour cell lines might provide a framework for interpreting the expression patterns in tumour specimens, we compared RNA isolated from two breast cancer biopsy samples, a sample of normal breast tissue and the NCI60 cell lines derived from breast cancers (excluding MDA-MB-435 and MDA-N) and leukaemias (Fig. 4). This clustering highlighted features of the gene expression pattern shared between the cancer specimens and individual cell lines derived from breast cancers and leukaemias.

The genes encoding keratin 8 (KRT8) and keratin 19 (KRT19), as well as most of the other 'epithelial' genes defined in the complete NCI60 cell line cluster, were expressed in both of the biopsy samples and the two breast-derived cell lines, MCF-7 and T47D, expressing the oestrogen receptor, suggesting that these transcripts originated in tumour cells with features similar to those of luminal epithelial cells (Fig. 5*a*). Expression of a set of genes characteristic of stromal cells, including collagen genes (*COL3A1*, *COL5A1* and *COL6A1*) and smooth muscle cell markers (*TAGLN*), was a feature shared by the tumour sample and the stromal-like cell lines Hs578T and BT549 (Fig. 5*b*). This feature of the expression pattern seen in the tumour samples is likely to be due to the stromal component of the tumour. The tumours also shared expression of a set of genes (Fig. 5*c*) with the multiple myeloma cell line (RPMI-8226), notably including immunoglobulin genes, consistent with the presence of B cells in the tumour (this was confirmed by staining with anti-

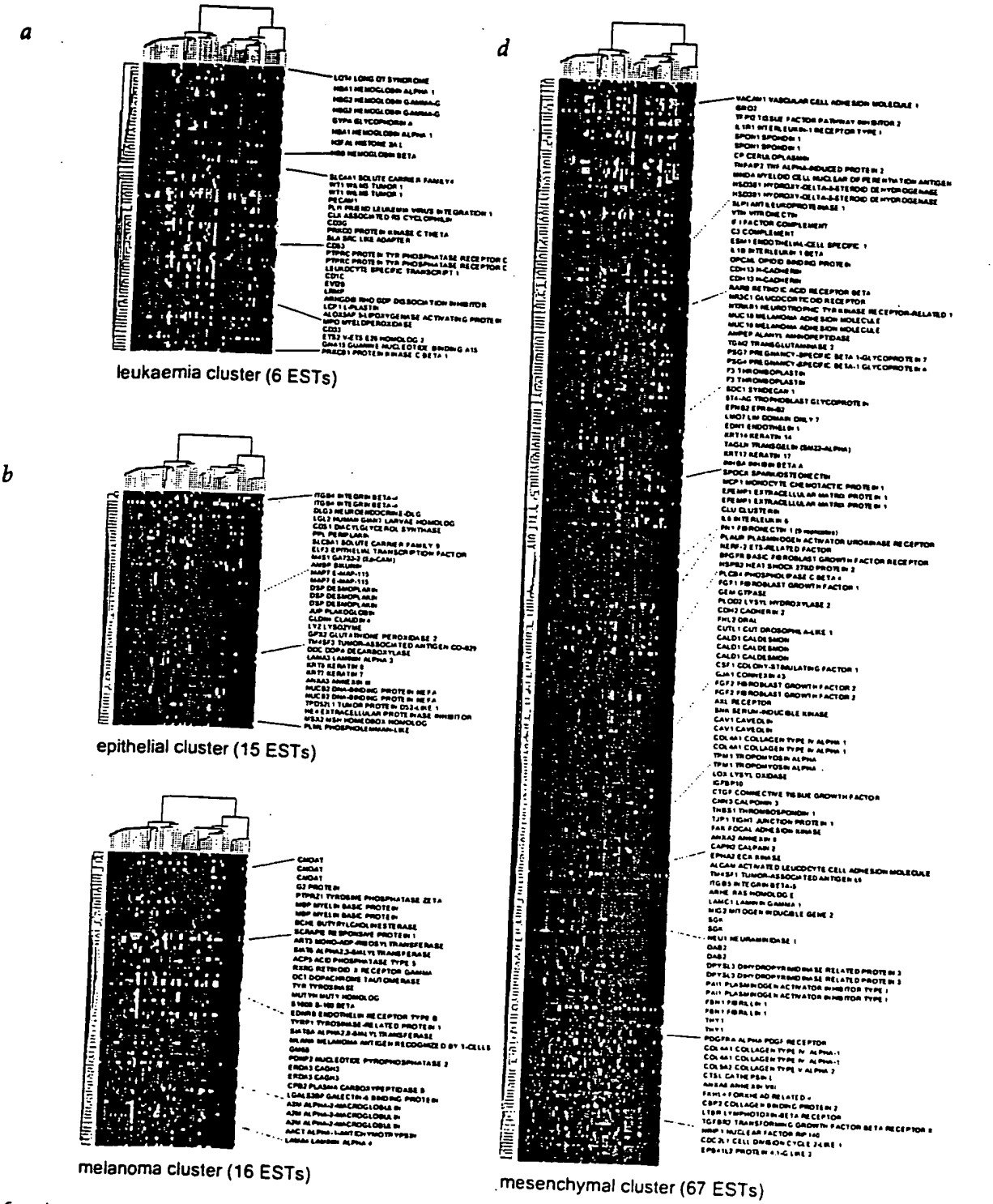

Fig. 3 Gene clusters related to tissue characteristics in the cell lines. Enlargements of the regions of the cluster diagram in Fig. 1 showing gene clusters enriched for genes expressed in cell lines of ostensibly similar origins. *a*, Cluster of genes highly expressed in the leukaemia-derived cell lines. Two sub-clusters distinguish genes that were expressed in most leukaemia-derived lines from those expressed exclusively in the eryroblastoid line, K562 (note that the triplicate hybridizations cluster together). *b*, Cluster of genes highly expressed in all colon (7/7) cell lines and all breast-derived cell lines positive for the oestrogen receptor (2/2). This set of genes was also moderately expressed in most ovarian lines (5/6) and some non-small-cell-lung (4/6) lines, but was expressed at a lower level in all renal-cancer–derived lines. *c*, Cluster of genes highly expressed in most melanoma-derived lines (6/7) and two related lines ostensibly derived from breast cancer (MDA-MB435 and MDA-N). *d*, Cluster of genes highly expressed in all glioblastoma (6/6) lines and most lines derived from renal-cell carcinoma (7/8), and more moderately expressed in a subset of carcinoma-derived lines. In all panels, names are shown only for all known genes whose identities were independently reverified by sequencing. The number of sequence-validated ESTs within the cluster is indicated below the cluster in parentheses. The position of gene names in the adjacent list only approximates their position in the cluster diagram as indicated by the lines connecting the colour chart with the gene list. Complete cluster images with all gene names and accession numbers are available (http://genome-www.stanford.edu/nci60).

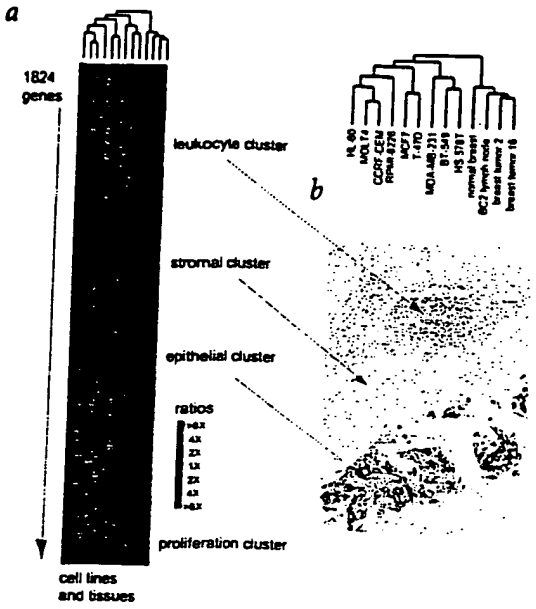

Fig. 4 Comparison of the gene expression patterns in clinical breast cancer specimens and cultured breast cancer and leukaemia cell lines. *a*, Two-dimensional hierarchical clustering applied to gene expression data for two breast cancer specimens, a lymph node metastasis from one patient, normal breast and the NCI60 breast and leukaemia-derived cell lines. The gene expression data from tissue specimens was clustered along with expression data from a subset of the NCI60 cell lines to explore whether features of expression patterns observed in specific lines could be identified in the tissue samples. Labels indicate gene clusters (shown in detail in Fig. 5) that may be related to specific cellular components of the tumour specimens. *b*, Breast cancer specimen 16 stained with anti-keratin antibodies, showing the complex mix of cell types characteristically found in breast tumours. The arrows highlight the different cellular components of this tissue specimen that were distinguished by the gene expression cluster analysis (Fig. 5).

immunoglobulin antibodies; data not shown). Therefore, distinct sets of genes with co-varying expression among the samples (Fig. 4, arrow) appear to represent distinct cell types that can be distinguished in breast cancer tissue. A fourth cluster of genes, more highly expressed in all of the cell lines than in any of the clinical specimens, was enriched for genes present in the 'proliferation' cluster described above (Fig. 5*d*). The variation in expression of these genes likely paralleled the difference in proliferation rate between the rapidly cycling cultured cell lines and the much more slowly dividing cells in tissues.

Discussion

Newly available genomics tools allowed us to explore variation in gene expression on a genomic scale in 60 cell lines derived from diverse tumour tissues. We used a simple cluster analysis to identify the prominent features in the gene expression patterns that appeared to reflect 'molecular signatures' of the tissue from which the cells originated. The histological characteristics of the cell lines that dominated the clustering were pervasive enough that similar relationships were revealed when alternative subsets of genes were selected for analysis. Additional features of the expression pattern may be related to variation in physiological attributes such as proliferation rate and activity of interferon-response pathways.

The properties of the tumour-derived cell lines in this study have presumably all been shaped by selection for resistance to host defences and chemotherapeutics and for rapid proliferation in the tissue culture environment of synthetic growth media, fetal bovine serum and a polystyrene substratum. But the primary identifiable factor accounting for variation in gene expression patterns among these 60 cell lines was the identity of the tissue from which each cell line was ostensibly derived. For most of the cell lines we examined, neither physiological nor experimental adaptation for growth in culture was sufficient to overwrite the gene expression programs established during differentiation *in vivo*. Nevertheless, the prominence of mesenchymal features in the cell lines isolated from glioblastomas and carcinomas may reflect a selection for the relative ease of establishment of cell lines expressing stromal characteristics, perhaps combined with physiological adaptation to tissue culture conditions[38–40].

Biological themes linking genes with related expression patterns may be inferred in many cases from the shared attributes of known genes within the clusters. Uncharacterized cDNAs are likely to encode proteins that have roles similar to those of the known gene products with which they appear to be co-regulated. Still, for several clusters of genes, we were unable to discern a common theme linking the identified members of the cluster. Further exploration of their variation in expression under more diverse conditions and more comprehensive investigation of the physiology of the NCI60 cells may provide insight[10]. The relationship of the gene expression patterns to the drug sensitivity patterns measured by the DTP is an example of linking variation in gene expression with more subtle and diverse phenotypic variation[11].

The patterns of gene expression measured in the NCI60 cell lines provide a framework that helps to distinguish the cells that express specific sets of genes in the histologically complex breast cancer specimens[41]. Although it is now feasible to analyse gene expression in micro-dissected tumour specimens[42,43], this observation suggests that it will be possible to explore and interpret some of the biology of clinical tumour samples by sampling them intact. As is useful in conventional morphological pathology, one might be able to observe interactions between a tumour and its microenvironment in this way. These relationships will be clarified by suitable analysis of gene expression patterns from intact as well as dissected tumours[12,14,15,41].

Methods

cDNA clones. We obtained the 9,703 human cDNA clones (Research Genetics) used in these experiments as bacterial colonies in 96-well microtitre plates[9]. Approximately 8,000 distinct Unigene clusters (representing nominally unique genes) were represented in this set of clones. All genes identified here by name represent clones whose identities were confirmed by re-sequencing, or by the criteria that two or more independent cDNA clones ostensibly representing the same gene had nearly identical gene expression patterns. A single-pass 3′ sequence re-verification was attempted for every clone after re-streaking for single colonies. For a subset of genes for which quality 3′ sequence was not obtained, we attempted to confirm identities by 5′ sequencing. Of the subset of clones selected for 5′ sequence verification on the basis of an interesting pattern of expression (888 total), 331 were correctly identified, 57, incorrectly identified, and 500, indeterminate (poor quality sequence). We estimated that 15%–20% of array elements contained DNA representing more than one clone per well. So far, the identities of ~3,000 clones have been verified. The full list of clones used and their nominal identities are available (gene names preceded by the designation "SID#" (Stanford Identification) represent clones whose identities have not yet been verified; http://genome-www.stanford.edu:8000/nci60).

Production of cDNA microarrays. The arrays used in this experiment were produced at Synteni Inc. (now Incyte Pharmaceuticals). Each insert was amplified from a bacterial colony by sampling 1 µl of bacterial media and performing PCR amplification of the insert using consensus primers for the three plasmids represented in the clone set (5′–TTGTAAAACGACG GCCAGTG–3′, 5′–CACACAGGAAACAGCTATG–3′). Each PCR product

(100 µl) was purified by gel exclusion, concentrated and resuspended in 3×SSC (10 µl). The PCR products were then printed on treated glass microscope slides using a robot with four printing tips. Detailed protocols for assembling and operating a microarray printer, and printing and experimental application of DNA microarrays are available (http://cmgm.stanford.edu/pbrown).

Preparation of mRNA and reference pool. Cell lines were grown from NCI DTP frozen stocks in RPMI-1640 supplemented with phenol red, glutamine (2 mM) and 5% fetal calf serum. To minimize the contribution of variations in culture conditions or cell density to differential gene expression, we grew each cell line to 80% confluence and isolated mRNA 24 h after transfer to fresh medium. The time between removal from the incubator and lysis of the cells in RNA stabilization buffer was minimized (<1 min). Cells were lysed in buffer containing guanidium isothiocyanate and total RNA was purified with the RNeasy purification kit (Qiagen). We purified mRNA as needed using a poly(A) purification kit (Oligotex, Qiagen) according to the manufacturer's instructions. Denaturing agarose gel electrophoresis assessed the integrity and relative contamination of mRNA with ribosomal RNA.

The breast tumours were surgically excised from patients and rapidly transported to the pathology laboratory, where samples for microarray analysis were quickly frozen in liquid nitrogen and stored at –80 °C until use. A frozen tumour specimen was removed from the freezer, cut into small pieces (~50–100 mg each), immediately placed into 10–12 ml of Trizol reagent (Gibco-BRL) and homogenized using a PowerGen 125 Tissue Homogenizer (Fisher Scientific), starting at 5,000 r.p.m. and gradually increasing to ~20,000 r.p.m. over a period of 30–60 s. We processed the Trizol/tumour homogenate as described in the Trizol protocol, including an initial step to remove fat. Once total RNA was obtained, we isolated mRNA with a FastTrack 2.0 kit (Invitrogen) using the manufacturer's protocol for isolating mRNA starting from total RNA. The normal breast samples were obtained from Clontech.

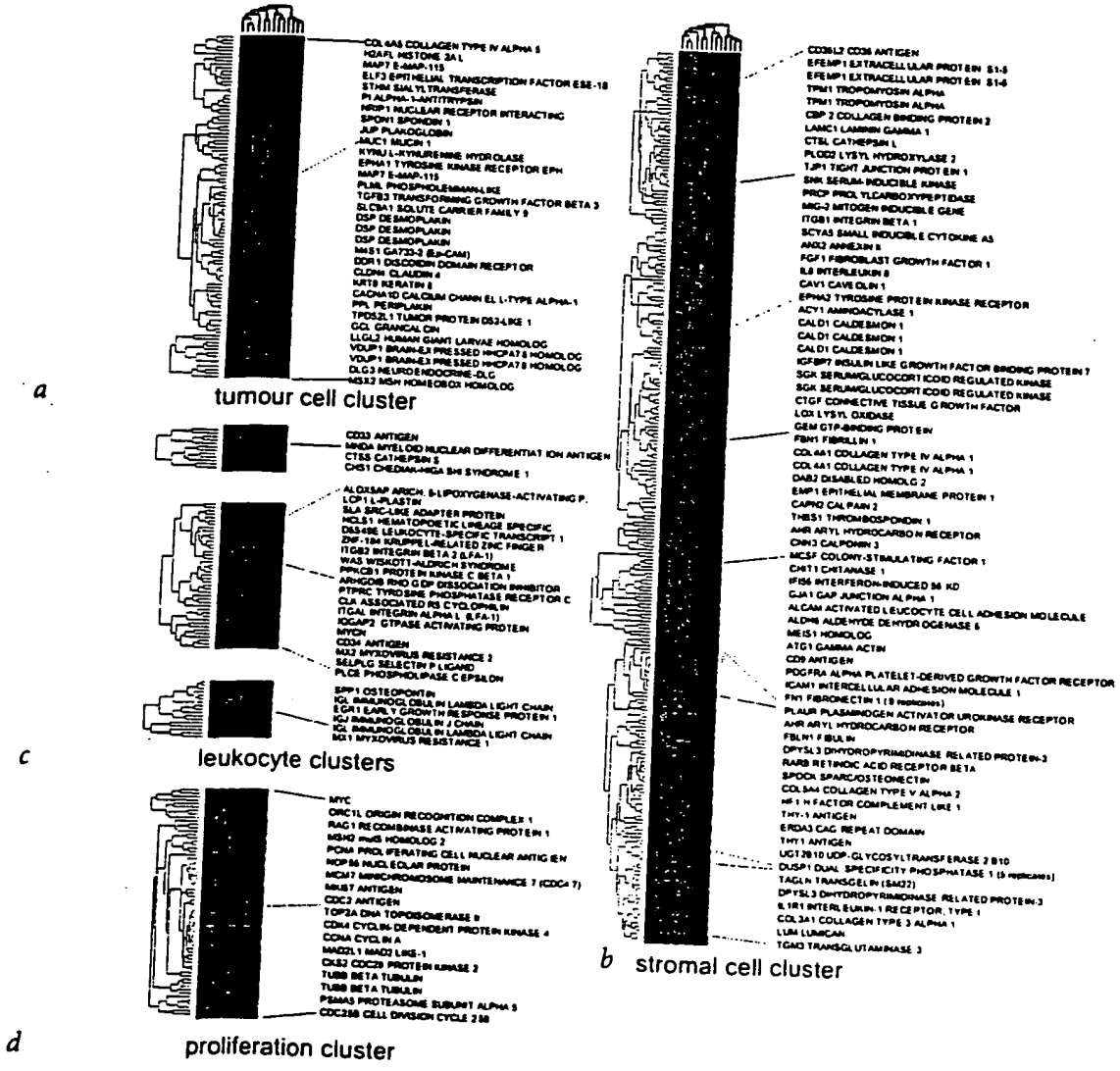

Fig. 5 Histologic features of breast cancer biopsies can be recognized and parsed based on gene expression patterns. Enlargements of the regions of the cluster diagram in Fig. 4 showing gene clusters enriched for genes expressed in different cell types in the breast cancer specimens, as distinguished by clustering with the cultured cell lines. *a*, A cluster including many genes characteristic of epithelial cells expressed in cell lines (T47D and MCF7) derived from breast cancer positive for the oestrogen receptor and tumours. *b*, Genes expressed in cell lines derived from breast cancer with stromal cell characteristics (Hs578T and BT549) and tumour specimens. Expression of these genes in the tumour samples may reflect the presence of myofibroblasts in the cancer specimen stroma. *c*, Genes expressed in leukocyte-derived cell lines, showing common leukocyte, and separate 'myeloid' and 'B-cell', gene clusters. *d*, Genes that were relatively highly expressed in all cell lines compared with the tumour specimens and normal breast. The higher expression of this set of genes involved in cell cycle transit in the cell lines is likely to reflect the higher proliferative rate of cells cultured in the presence of serum compared with the average proliferation rate of cells in the biopsied tissue.

We combined mRNA from the following cells in equal quantities to make the reference pool: HL-60 (acute myeloid leukaemia) and K562 (chronic myeloid leukaemia); NCI-H226 (non-small-cell-lung); COLO 205 (colon); SNB-19 (central nervous system); LOX-IMVI (melanoma); OVCAR-3 and OVCAR-4 (ovarian); CAKI-1 (renal); PC-3 (prostate); and MCF7 and Hs578T (breast). The criterion for selection of the cell lines in the reference are described in detail in the accompanying manuscript[12].

Doubling-time calculations. We calculated doubling times based on routine NCI60 cell line compound screening data; and they reflect the doubling times for cells inoculated into 96-well plates at the screening inoculation densities and grown in RPMI 1640 medium supplemented with 5% fetal bovine serum for 48 h. We measured cell populations using sulforhodamine B optical density measurement assay. The doubling time constant k was calculated using the equation: $N/No = e^{kt}$, where No is optical density for control (untreated) cells at time zero, N is optical density for control cells after 48-h incubation, and t is 48 h. The same equation was then used with the derived k to calculate the doubling time t by setting N/No = 2. For a given cell line, we obtained No and N values by averaging optical densities (N>6,000) obtained for each cell line for a year's screening. Data and experimental details are available (http://dtp.nci.nih.gov).

Preparation and hybridization of fluorescent labelled cDNA. For each comparative array hybridization, labelled cDNA was synthesized by reverse transcription from test cell mRNA in the presence of Cy5-dUTP, and from the reference mRNA with Cy3-dUTP, using the Superscript II reverse-transcription kit (Gibco-BRL). For each reverse transcription reaction, mRNA (2 μg) was mixed with an anchored oligo-dT (d-20T-d(AGC)) primer (4 μg) in a total volume of 15 μl, heated to 70 °C for 10 min and cooled on ice. To this sample, we added an unlabelled nucleotide pool (0.6 μl; 25 mM each dATP, dCTP, dGTP, and 15 mM dTTP), either Cy3 or Cy5 conjugated dUTP (3 μl; 1 mM; Amersham), 5×first-strand buffer (6 μl; 250 mM Tris-HCL, pH 8.3, 375 mM KCl, 15 mM $MgCl_2$), 0.1 M DTT (3 μl) and 2 μl of Superscript II reverse transcriptase (200 μ/μl). After a 2-h incubation at 42 °C, the RNA was degraded by adding 1 N NaOH (1.5 μl) and incubating at 70 °C for 10 min. The mixture was neutralized by adding of 1 N HCL (1.5 μl), and the volume brought to 500 μl with TE (10 mM Tris, 1 mM EDTA). We added Cot1 human DNA (20 μg; Gibco-BRL), and purified the probe by centrifugation in a Centricon-30 micro-concentrator (Amicon). The two separate probes were combined, brought to a volume of 500 μl, and concentrated again to a volume of less than 7 μl. We added 10 μg/μl poly(A) RNA (1 μl; Sigma) and tRNA (10 μg/μl; Gibco-BRL) were added, and adjusted the volume to 9.5 μl with distilled water. For final probe preparation, 20×SSC (2.1 μl; 1.5 M NaCl, 150 mM NaCitrate, pH 8.0) and 10% SDS (0.35 μl) were added to a total final volume of 12 μl. The probes were denatured by heating for 2 min at 100 °C, incubated at 37 °C for 20–30 min, and placed on the array under a 22 mm×22 mm glass coverslip. We incubated slides overnight at 65 °C for 14–18 h in a custom slide chamber with humidity maintained by a small reservoir of 3×SSC. Arrays were washed by submersion and agitation for 2–5 min in 2×SSC with 0.1% SDS, followed by 1×SSC and then 0.1×SSC. The arrays were "spun dry" by centrifugation for 2 min in a slide-rack in a Beckman GS-6 tabletop centrifuge in Microplus carriers at 650 r.p.m. for 2 min.

Array quantitation and data processing. Following hybridization, arrays were scanned using a laser-scanning microscope (ref. 17; http://cmgm.stanford.edu/pbrown). Separate images were acquired for Cy3 and Cy5. We carried out data reduction with the program ScanAlyze (M.B.E., available at http://rana.stanford.edu/software). Each spot was defined by manual positioning of a grid of circles over the array image. For each fluorescent image, the average pixel intensity within each circle was determined, and a local background was computed for each spot equal to the median pixel intensity in a square of 40 pixels in width and height centred on the spot centre, excluding all pixels within any defined spots. Net signal was determined by subtraction of this local background from the average intensity for each spot. Spots deemed unsuitable for accurate quantitation because of array artefacts were manually flagged and excluded from further analysis. Data files generated by ScanAlyze were entered into a custom database that maintains web-accessible files. Signal intensities between the two fluorescent images were normalized by applying a uniform scale factor to all intensities measured for the Cy5 channel. The normalization factor was chosen so that the mean log(Cy3/Cy5) for a subset of spots that achieved a minimum quality parameter (approximately 6,000 spots) was 0. This effectively defined the signal-intensity-weighted 'average' spot on each array to have a Cy3/Cy5 ratio of 1.0.

Cluster analysis. We extracted tables (rows of genes, columns of individual microarray hybridizations) of normalized fluorescence ratios from the database. Various selection criteria, discussed in relation to each data set, were applied to select subsets of genes from the 9,703 cDNA elements on the arrays. Before clustering and display, the logarithm of the measured fluorescence ratios for each gene were centred by subtracting the arithmetic mean of all ratios measured for that gene. The centring makes all subsequent analyses independent of the amount of each gene's mRNA in the reference pool.

We applied a hierarchical clustering algorithm separately to the cell lines and genes using the Pearson correlation coefficient as the measure of similarity and average linkage clustering[3,19–21]. The results of this process are two dendrograms (trees), one for the cell lines and one for the genes, in which very similar elements are connected by short branches, and longer branches join elements with diminishing degrees of similarity. For visual display the rows and columns in the initial data table were reordered to conform to the structures of the dendrograms obtained from the cluster analysis. Each cell in the cluster-ordered data table was replaced by a graded colour (pure red through black to pure green), representing the mean-adjusted ratio value in the cell. Gene labels in cluster diagrams are displayed here only for genes that were represented in the microarray by sequence-verified cDNAs. A complete software implementation of this process is available (http://rana.stanford.edu/software), as well as all clustering results (http://genome-www.stanford.edu/nci60).

Acknowledgements

We thank members of the Brown and Botstein labs for helpful discussions. This work was supported by the Howard Hughes Medical Institute and a grant from the National Cancer Institute (CA 077097). The work of U.S. and J.N.W. was supported in part by a grant from the National Cancer Institute Breast Cancer Think Tank. D.T.R is a Walter and Idun Berry Fellow. M.B.E. is an Alfred P. Sloan Foundation Fellow in Computational Molecular Biology. C.M.P. is a SmithKline Beecham Pharmaceuticals Fellow of the Life Science Research Foundation. P.O.B. is an Associate Investigator of the Howard Hughes Medical Institute.

Received 20 July 1999; accepted 13 January 2000.

1. Stinson, S.F. *et al.* Morphological and immunocytochemical characteristics of human tumor cell lines for use in a disease-oriented anticancer drug screen. *Anticancer Res.* **12**, 1035–1053 (1992).
2. Myers, T.G. *et al.* A protein expression database for the molecular pharmacology of cancer. *Electrophoresis* **18**, 647–653 (1997).
3. Weinstein, J.N. *et al.* An information-intensive approach to the molecular pharmacology of cancer. *Science* **275**, 343–349 (1997).
4. Monks, A., Scudiero, D.A., Johnson, G.S., Paull, K.D. & Sausville, E.A. The NCI anti-cancer drug screen: a smart screen to identify effectors of novel targets. *Anticancer Drug Des.* **12**, 533–541 (1997).
5. Paull, K.D. *et al.* Display and analysis of patterns of differential activity of drugs against human tumor cell lines: development of mean graph and COMPARE algorithm. *J. Natl Cancer Inst.* **81**, 1088–1092 (1989).
6. Weinstein, J.N. *et al.* Neural computing in cancer drug development: predicting mechanism of action. *Science* **258**, 447–451 (1992).
7. van Osdol, W.W., Myers, T.G., Paull, K.D., Kohn, K.W. & Weinstein, J.N. Use of the Kohonen self-organizing map to study the mechanisms of action of chemotherapeutic agents. *J. Natl Cancer Inst.* **86**, 1853–1859 (1994).
8. DeRisi, J.L., Iyer, V.R. & Brown, P.O. Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science* **278**, 680–686 (1997).
9. Iyer, V.R. *et al.* The transcriptional program in the response of human fibroblasts to serum. *Science* **283**, 83–87 (1999).
10. Brown, P.O. & Botstein, D. Exploring the new world of the genome with DNA microarrays. *Nature Genet.* **21** (suppl.), 33–37 (1999).
11. Scherf, U. *et al.* A gene expression database for the molecular pharmacology of cancer. *Nature Genet.* **24**, 236–244 (2000).
12. Khan, J. *et al.* Gene expression profiling of alveolar rhabdomyosarcoma with cDNA microarrays. *Cancer Res.* **58**, 5009–5013 (1998).
13. Der, S.D., Zhou, A., Williams, B.R. & Silverman, R.H. Identification of genes differentially regulated by interferon-α, -β or -γ or using oligonucleotide arrays. *Proc. Natl Acad. Sci. USA* **95**, 15623–15628 (1998).
14. Alon, U. *et al.* Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc. Natl Acad. Sci. USA* **96**, 6745–6750 (1999).
15. Wang, K. *et al.* Monitoring gene expression profile changes in ovarian carcinomas using cDNA microarray. *Gene* **229**, 101–108 (1999).
16. Tamayo, P. *et al.* Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *Proc. Natl Acad. Sci. USA* **96**, 2907–2912 (1999).
17. Shalon, D., Smith, S.J. & Brown, P.O. A DNA microarray system for analyzing complex DNA samples using two-color fluorescent probe hybridization. *Genome Res.* **6**, 639–645 (1996).
18. Eisen, M.B. & Brown, P.O. DNA arrays for analysis of gene expression. *Methods Enzymol.* **303**, 179–205 (1999).
19. Sokal, R.R. & Sneath, P.H.A. *Principles of Numerical Taxonomy* (W.H. Freeman, San Francisco, 1963).
20. Hartigan, J.A. *Clustering Algorithms* (Wiley, New York, 1975).
21. Eisen, M.B., Spellman, P.T., Brown, P.O. & Botstein, D. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl Acad. Sci. USA* **95**, 14863–14868 (1998).
22. del Marmol, V. & Beermann, F. Tyrosinase and related proteins in mammalian pigmentation. *FEBS Lett.* **381**, 165–168 (1996).
23. Kawakami, Y. *et al.* The use of melanosomal proteins in the immunotherapy of melanoma. *J. Immunother.* **21**, 237–246 (1998).
24. Cailleau, R., Olive, M. & Cruciger, Q.V. Long-term human breast carcinoma cell lines of metastatic origin: preliminary characterization. *In Vitro* **14**, 911–915 (1978).
25. Brinkley, B.R. *et al.* Variations in cell form and cytoskeleton in human breast carcinoma cells in vitro. *Cancer Res.* **40**, 3118–3129 (1980).
26. Nesland, J.M., Holm, R., Johannessen, J.V. & Gould, V.E. Neuroendocrine differentiation in breast lesions. *Pathol. Res. Pract.* **183**, 214–221 (1988).
27. Davies, J.A. & Garrod, D.R. Molecular aspects of the epithelial phenotype. *Bioessays* **19**, 699–704 (1997).
28. Garrod, D., Chidgey, M. & North, A. Desmosomes: differentiation, development, dynamics and disease. *Curr. Opin. Cell Biol.* **8**, 670–678 (1996).
29. Cowin, P. & Burke, B. Cytoskeleton-membrane interactions. *Curr. Opin. Cell Biol.* **8**, 56–65 (1996); erratum: **8**, 244 (1996).
30. Litvinov, S.V. *et al.* Epithelial cell adhesion molecule (Ep-CAM) modulates cell-cell interactions mediated by classic cadherins. *J. Cell Biol.* **139**, 1337–1348 (1997).
31. Helmle-Kolb, C. *et al.* Na/H exchange activities in NHE1-transfected OK-cells: cell polarity and regulation. *Pflugers Arch.* **425**, 34–40 (1993); erratum: **427**, 387 (1994).
32. Manfruelli, P., Arquier, N., Hanratty, W.P. & Semeriva, M. The tumor suppressor gene, lethal(2)giant larvae (l(2)g1), is required for cell shape change of epithelial cells during Drosophila development. *Development* **122**, 2283–2294 (1996).
33. Lincecum, J.M., Fannon, A., Song, K., Wang, Y. & Sassoon, D.A. Msh homeobox genes regulate cadherin-mediated cell adhesion and cell-cell sorting. *J. Cell Biochem.* **70**, 22–28 (1998).
34. Hackett, A.J. *et al.* Two syngeneic cell lines from human breast tissue: the aneuploid mammary epithelial (Hs578T) and the diploid myoepithelial (Hs578Bst) cell lines. *J. Natl Cancer Inst.* **58**, 1795–1806 (1977).
35. Rutka, J.T. *et al.* Establishment and characterization of a cell line from a human gliosarcoma. *Cancer Res.* **46**, 5893–5902 (1986).
36. Nguyen, H., Hiscott, J. & Pitha, P.M. The growing family of interferon regulatory factors. *Cytokine Growth Factor Rev.* **8**, 293–312 (1997).
37. Moscow, J.A., Schneider, E., Ivy, S.P. & Cowan, K.H. Multidrug resistance. *Cancer Chemother. Biol. Response Modif.* **17**, 139–177 (1997).
38. Smith, H.S. & Hackett, A.J. The use of cultured human mammary epithelial cells in defining malignant progression. *Ann. N Y Acad. Sci.* **464**, 288–300 (1986).
39. Rutka, J.T. *et al.* Establishment and characterization of five cell lines derived from human malignant gliomas. *Acta Neuropathol.* **75**, 92–103 (1987).
40. Ronnov-Jessen, L., Petersen, O.W. & Bissell, M.J. Cellular changes involved in conversion of normal to malignant breast: importance of the stromal reaction. *Physiol. Rev.* **76**, 69–125 (1996).
41. Perou, C.M. *et al.* Distinctive gene expression patterns in human mammary epithelial cells and breast cancers. *Proc. Natl Acad. Sci. USA* **96**, 9212–9217 (1999).
42. Bonner, R.F. *et al.* Laser capture microdissection: molecular analysis of tissue. *Science* **278**, 1481–1483 (1997).
43. Sgroi, D.C. *et al.* In vivo gene expression profile analysis of human breast cancer progression. *Cancer Res.* **59**, 5656–5661 (1999).